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

Supplementary Material

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
jiahuanc@foxmail.com, ly10061105@gmail.com, yongxin_shi@foxmail.com
danny_ding@intsig.net, eelwjin@scut.edu.cn

A Datasheet for WenMind

A.1 Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

A1: The purpose of WenMind is to quantify the capabilities of existing large language models (LLMs) on Chinese classical literature and language arts (CCLLA), providing a reference for the development of this field. Existing benchmarks focus primarily on the Ancient Prose sub-domain or contain only a limited number of tasks, making it challenging to provide a thorough and holistic assessment of LLMs' capabilities in CCLLA. To fill this gap, we propose WenMind, a comprehensive benchmark dedicated for evaluating LLMs in CCLLA. WenMind covers the sub-domains of Ancient Prose, Ancient Poetry, and Ancient Literary Culture, comprising 4,875 question-answer pairs, spanning 42 fine-grained tasks, 3 question formats, and 2 evaluation scenarios: domain-oriented and capability-oriented. Overall, WenMind serves as a standardized and comprehensive baseline, providing valuable insights for future CCLLA research.

2. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

A2: The WenMind benchmark is created by the Deep Learning and Vision Computing Lab (DLVC-Lab) of South China University of Technology, INTSIG Information Co., Ltd, and INTSIG-SCUT Joint Lab on Document Analysis and Recognition.

A.2 Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

A1: The Wenmind benchmark consists of 4,875 plain text questions. Each question is represented in a dictionary format, containing the following keys: ID, domain, capability, question format,

[†]Equal contribution
[*]Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

coarse-grained task, fine-grained task, question, and answer. The dataset is entirely stored in a JSON file.

2. How many instances are there in total (of each type, if appropriate)?

A2: The WenMind benchmark comprises a total of 4,875 instances, which are categorized into three distinct sections: 1,900 instances in the Ancient Prose section, 1,845 instances in the Ancient Poetry section, and 1,130 instances in the Ancient Literature Culture section.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this represent ativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances because instances were withheld or unavailable).

A3: The WenMind benchmark encompasses all possible instances.

4. What data does each instance consist of?

A4: The WenMind benchmark consists of textual data. Each text instance includes the following information: ID, domain, capability, question format, coarse-grained task, fine-grained task, question, and answer.

5. Is there a label or target associated with each instance?

A5: The WenMind benchmark includes questions and their corresponding answers, along with additional information such as the question format and task name.

6. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

A6: No, each of our data instances contains comprehensive information.

7. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

A7: No, our data instances are independent of each other and have no interrelations.

8. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

A8: No, our data is utilized solely for the purpose of testing and not for training.

9. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

A9: In WenMind, we have incorporated extensive manual calibration to minimize errors and noise as much as possible, yet we cannot guarantee that every answer will be flawless.

10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

A10: Yes, the WenMind benchmark is self-contained.

11. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

A11: No, we construct the WenMind benchmark from publicly accessible sources.

12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

A12: No, the WenMind benchmark does not contain any content that is offensive, insulting, threatening, or might otherwise cause anxiety.

13. Does the dataset relate to people?

A13: No, the WenMind benchmark has nothing to do with people.

A.3 Collection Process

1. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

A1: The data construction and specific sources are thoroughly described in Section 3.2 and Appendix B.3. The text we gather from the Internet is sourced from several open copyright websites, such as `https://www.sou-yun.cn/` and `https://www.zdic.net/`.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

A2: Our data collection is carried out through three mechanisms: manual human curation, a web crawling program, and the API of LLMs.

3. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)

A3: The WenMind benchmark is created by the researchers from Deep Learning and Vision Computing Lab (DLVC-Lab) of South China University of Technology, INTSIG Information Co., Ltd, and INTSIG-SCUT Joint Lab on Document Analysis and Recognition.

4. Over what timeframe was the data collected?

A4: Our data collection spans a period of two months.

5. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

A5: We mainly collect the data from the Internet, open-source datasets, and LLM generation.

A.4 Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

A1: Yes, the detail is discussed in Section 3.2.

2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

A2: No, the data we provide has undergone processing and transformation, combining original texts with annotations for storage.

A.5 Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.

A1: No, the WenMind benchmark has not appeared in previous works.

2. What (other) tasks could the dataset be used for?

A2: No, the WenMind benchmark is solely used for the evaluation of LLMs.

3. Are there tasks for which the dataset should not be used? If so, please provide a description.

A3: To prevent data leakage and ensure the fairness of the evaluation [1, 2], the WenMind benchmark cannot be used for training LLMs.

A.6 Distribution

1. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

A1: Our dataset is publicly available at https://github.com/SCUT-DLVCLab/WenMind.

2. When will the dataset be distributed?

A2: Our dataset has been already distributed.

3. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

A3: For details, please refer to Appendix F.

A.7 Maintenance

1. Who will be supporting/hosting/maintaining the dataset?

A1: The Deep Learning and Vision Computing Lab (DLVC-Lab) of South China University of Technology.

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

A2: The manager can be contacted through the email address or Github.

3. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

A3: We will publish a correction list for the WenMind benchmark on GitHub every quarter.

4. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

A4: We will maintain all versions of the WenMind benchmark on GitHub.

5. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

A5: If other researchers or individuals are interested in extending, augmenting, building on, or contributing to the dataset, they should contact us via email, clearly articulating their intentions and requesting our consent prior to any further actions.

B Dataset

B.1 Task Description

Table 1 provides detailed information for 42 fine-grained tasks, including the Chinese and English task names, task descriptions, assessed capabilities, scale, average question length, average answer length, and question format.

Table 1: Detail of each task in WenMind. “Q” represents “Question”; “A” represents “Answer”; “FB” represents “Fill-in-the-Blank questions”; “MCQ” represents “Multiple-Choice Questions”; “QA” represents “Question-and-Answer questions”. Zoom in for better view.

ID	Task Name	Task Description	Capability	Scale	Avg.Q Tokens	Avg.A Tokens	Question Format		
							FB	MCQ	QA
T1-1	Inverted Sentence Structure	Correct word order for inverted sentences	Understanding	18	22.22	8.89			4
T1-2	Elliptical Sentence	Answer the omitted information in the elliptical sentence	Understanding	32	34.88	33.25			4
T1-3	Inverted Sentence Types	Identify the inversion type of inverted sentences	Understanding	7	55.43	1.00		4	
T1-4	Sentence Structure Identification	Identify the sentence’s syntactic type	Understanding	43	46.33	1.00		4	
T2	Classical Chinese to Modern Chinese	Translate classical Chinese into modern Chinese	Understanding	200	31.91	30.74			4
T3	Modern Chinese to Classical Chinese	Translate modern Chinese into classical Chinese	Understanding	200	42.46	19.56			4
T4	Named Entity Recognition	Extract named entities from ancient prose sentences	Understanding	200	236.99	43.20			4
T5	Punctuation	Add punctuation to ancient prose sentences	Understanding	200	34.95	26.50			4
T6	Topic Classification	Select theme categories based on ancient prose sentences	Understanding	200	172.40	2.00			4
T7	Word Explanation	Explain the words and phrases in ancient prose sentences	Understanding	100	32.84	27.21			4
T8	Reading Comprehension	Read ancient prose texts and answer related questions	Understanding	100	207.04	81.47			4
T9	Function Words	Answer the usage of function words in ancient prose sentences	Understanding	100	76.88	1.00		4	
T10	Homophones	Identify whether a character is a homophone	Understanding	200	145.94	1.00		4	
T11	Polysemy	Distinguish between different meanings of the same character	Understanding	200	127.40	1.00		4	
T12	Ancient Prose Writing	Writing in classical Chinese	Generation	100	102.96	809.96			4
T13-1	Appreciation Exam Questions	Answer appreciation questions based on ancient poetry	Understanding	150	200.69	68.42		4	4
T13-2	Free Appreciation	Conduct a free and detailed analysis of ancient poetry	Understanding	100	112.24	189.24			4
T14-1	Poetry Writing	Compose a poem based on the theme	Generation	30	18.97	59.83			4
T14-2	Ci Writing	Compose a Ci based on the theme	Generation	50	25.14	84.98			4
T14-3	Qu Writing	Compose a Qu based on the theme	Generation	20	25.15	54.40			4
T15-1	Content Q&A	Answer the complete content of ancient poetry according to the title and author	Knowledge	200	29.62	59.88			4
T15-2	Title and Author Q&A	Answer the title and author according to the content of ancient poetry	Knowledge	200	85.25	15.76			4
T15-3	Write the Next Sentence	Write the next sentence according to the previous sentence in the ancient poem	Knowledge	100	34.67	6.36			4
T15-4	Write the Previous Sentence	Write the previous sentence according to the next sentence in the ancient poem	Knowledge	100	35.83	6.24			4
T15-5	Comprehension Dictation	Provide ancient poetry sentences that meet the requirements	Knowledge	30	67.93	14.93	4		
T15-6	Genre Judgment	Judge the genre of ancient poetry	Knowledge	120	87.53	3.25			4
T16	Ancient Poetry Translation	Translate ancient poetry into modern Chinese	Understanding	200	80.71	104.61			4
T17	Sentiment Classification	Judge the sentiment contained in ancient poetry	Understanding	200	111.16	3.28		4	
T18	Ancient Poetry to English	Translate ancient poetry into English	Understanding	50	73.24	392.98			4
T19	Poet Introduction	Provide a detailed introduction of the poet	Knowledge	110	11.53	346.51			4
T20	Analysis of Imagery	Provide the meanings of the imagery	Knowledge	185	32.03	228.39			4
T21-1	Couplet Following	Create the following couplet based on the previous one	Generation	100	19.92	9.92			4
T21-2	Couplet Writing	Write a couplet based on the theme	Generation	100	16.48	18.78			4
T21-3	HengPi Writing	Write HengPi based on the content of a couplet	Generation	100	28.80	4.00			4
T22-1	Synonyms	Provide the synonym for the idiom	Knowledge	100	16.19	6.40			4
T22-2	The Origin of Idiom	Provide the source of the idiom	Knowledge	100	13.33	41.40			4
T22-3	Idiom Finding	Extract idioms from ancient Chinese sentences and provide their meanings	Knowledge	100	66.45	30.10			4
T22-4	Idiom Explanation	Provide the meaning of idioms	Knowledge	100	15.14	28.46			4
T23	Riddle	Guess the answer based on clues or clever hints	Knowledge	100	17.94	2.14			4
T24	Xiehouyu	Complete the second half of the proverb based on the first half	Knowledge	100	16.52	5.06			4
T25	Historical Chinese Phonology	Answer questions about ancient Chinese phonetics and rhymes	Knowledge	100	61.99	1.00		4	
T26	Knowledge of Sinology Q&A	Answer questions about Sinology	Knowledge	130	22.12	10.69		4	4

B.2 Task Example

Figure 1, 2, 3, 4, 5 and 6 respectively present examples of question-answer pairs for the 42 fine-grained tasks.

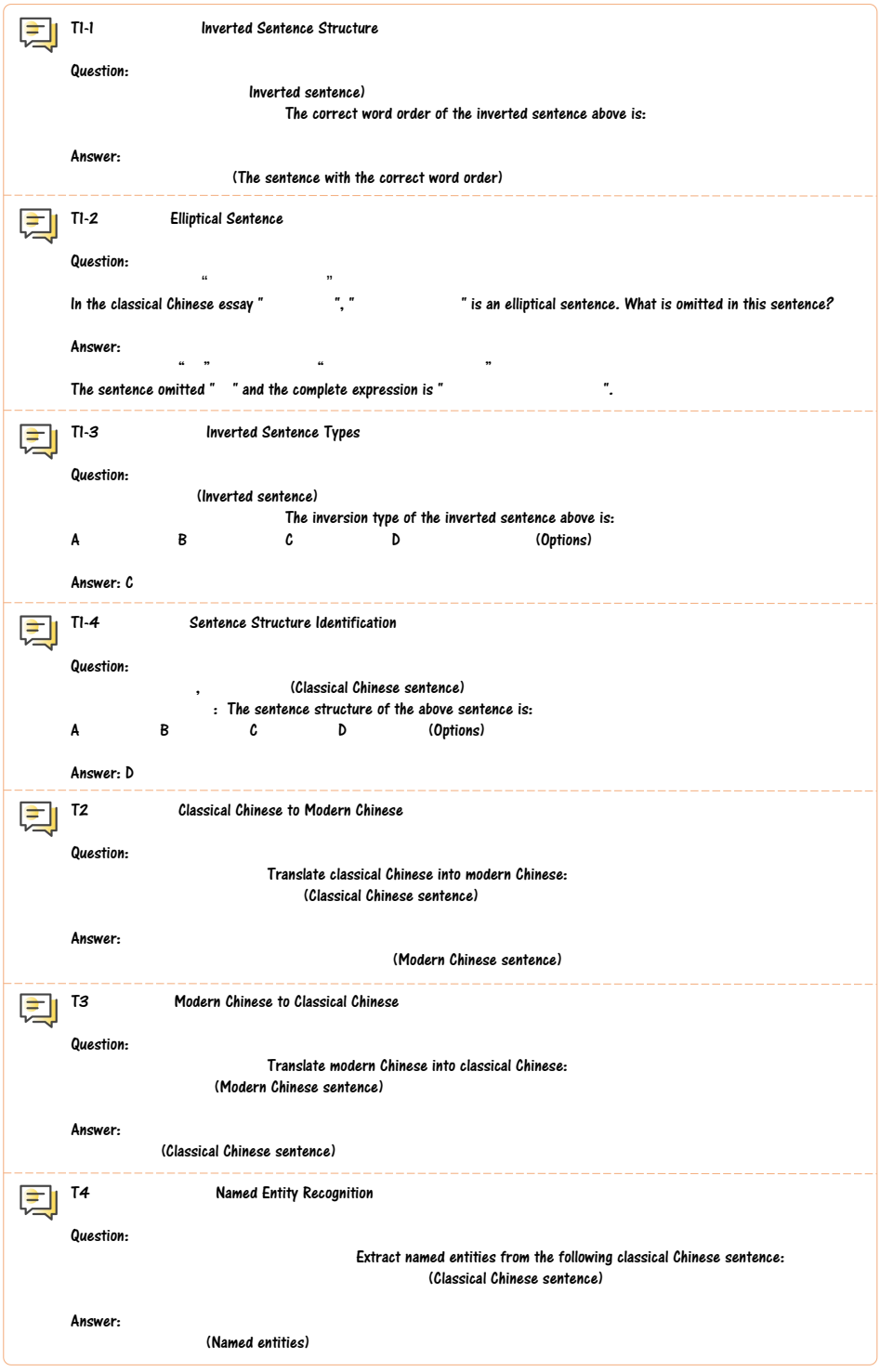

Figure 1: Task 1-1 to Task 4 examples.

T5 Punctuation

Question:

Add punctuation to the following sentence:

(Classical Chinese sentence)

Answer:

(Classical Chinese sentence with punctuation added)

T6 Topic Classification

Question:

Categories

What category does the topic of the following sentence belong to:

(Classical Chiese sentence)

Answer:

(Category)

T7 Word Explanation

Question:

“ ” “ ”

What is the meaning of the character " " in the classical Chinese sentence ' '?

Answer:

“ ” “ ”

Here, the character " " means "exchange", referring to trading one thing for another.

T8 Reading Comprehension

Question:

Please answer the question based on classical Chinese.

..... (Classical Chinese sentences)

How did Pei Lue finally obtain an official position?

Answer:

Pei Lue responded to Wen Yanbo's test with a sharp mockery by praising himself and showing his talent, which ultimately made Wen Yanbo feel ashamed and earned him an official position.

T9 Function Words

Question:

(Classical Chinese sentence)

“ ” The usage of " " in the above sentence is:

A B C D (Options)

Answer: B

T10 Homophones

Question:

[] “ ”

In the following options, which of the " " in the brackets [] is a homophone:

A [ ] B [ ] C [ ] D [ ] (Options)

Answer: C

T11 Polysemy

Question:

[] “ ” “ ”

The character " " in the brackets [] in the following options is translated as "similar, ally" is

A [ ] B [ ]

C [ ] D [ ] (Options)

Answer: A

Figure 2: Task 5 to Task 11 examples.

T12 Ancient Prose Writing

Question:

Please help me write an enrollment advertisement for a weight loss camp, where no fees are charged if the weight is not lost, in classical Chinese.

Answer:

“ ”

..... (The content written in classical Chinese)

T13-1 Appreciation Exam Questions

Question:

.

(Ancient poem content)

What is the main theme this poem aims to convey?

Answer:

Parting sorrow and separation grief (or homesickness)

T13-2 Free Appreciation

Question:

.

(Ancient poem content)

Please provide a free appreciation of this ancient poem.

Answer:

The poem is titled Spring Sorrow, which is written with deep and affectionate expression, showing the heroine's persistent and innocent emotions.

T14-1 Poetry Writing

Question:

“ ” Please create a five-character quatrain with the theme of "summer".

Answer:

(Ancient poem content)

T14-2 Ci Writing

Question:

“ ”

Please create a ci with the theme of "zither and lute", using the ci title " ".

Answer:

.

(Ci content)

T14-3 Qu Writing

Question:

“ ”

Please create a qu with the theme of "artificial intelligence", using the qu title " ".

Answer:

.

(Qu content)

T15-1 Content Q&A

Question:

Please provide the full text of the poem "Ascending Feilai Peak" by the poet Wang Anshi (Song Dynasty).

Answer:

(Ancient poem content)

Figure 3: Task 12 to Task 15-1 examples.

T15-2 Title and Author Q&A

Question:

Based on the content of the ancient poem provided below, provide the corresponding title and author.
(Ancient poem content)

Answer:
(The title and author)

T15-3 Write the Next Sentence

Question:
“ ”
The next sentence after the ancient poetry phrase " " written by Li Bai in " " is what?

Answer:
(The content of the next sentence)

T15-4 Write the Previous Sentence

Question:
“ ”
The previous sentence before the ancient poetry phrase " " written by Wang Wei in " " is what?

Answer:
(The content of the previous sentence)

T15-5 Comprehension Dictation

Question:
_____ _____
The main theme sentence of " " is _____ _____

Answer:
(The main theme sentence)

T15-6 Genre Judgment

Question:
(Ancient poem content)
What is the genre of this ancient Chinese poem?

Answer:
Five character quatrain

T16 Ancient Poetry Translation

Question:
Please translate the following classical Chinese poem into modern Chinese.

(Ancient poem content)

Answer:

(Modern Chinese sentence)

T17 Sentiment Classification

Question:
(Ancient poem content)
“ ” “ ” “ ” “ ” “ ”

Please select the most suitable option among "negative", "implicitly negative", "neutral", "implicitly positive", and "positive" (with negative emotions ranging from strong to weak) to describe the sentiment of the above ancient poem.

Answer:
(Positive)

Figure 4: Task 15-2 to Task 17 examples.

T18 Ancient Poetry to English

Question:

Please translate the content of the following ancient poem into English:
(Ancient poem content)

Answer:
Rain is pitterpattering outside the window. Springtime spirit is on the wane.

T19 Poet Introduction

Question:

“ ” Please introduce the poet "Su Shi".

Answer:
1037 —1101 .....
(Information about Su Shi)

T20 Analysis of Imagery

Question:
“ ” /
" " appears as an image in ancient Chinese poetry and prose. What are the common meanings associated with this image?

Answer:
(1) (2)
(1) Blue silk fabric. (2) A metaphor for the winding and clear blue water.

T21-1 Couplet Following

Question:

Match the couplet according to the first line:
(The first line content)

Answer:

(The second line content)

T21-2 Couplet Writing

Question:

Creating a couplet for the following festival: Lantern Festival

Answer:

(The content of the couplet)

T21-3 HengPi Writing

Question:

Create a HengPi for the following couplet:
(The content of the couplet)

Answer:
(The content of the HengPi)

T22-1 Synonyms

Question:

Provide a synonym for the following idiom:
(Idiom)

Answer:

(Synonyms)

Figure 5: Task 18 to Task 22-1 examples.

T22-2 The Origin of Idiom

Question:

Where does the idiom " " originate from?

Answer:
. “ ”
(The origin of idiom)

T22-3 Idiom Finding
Question:
▪ “ ” (Ancient poem content)

From the content above, find the idiom hidden within and explain its meaning.

Answer:

(The idiom and its meaning)

T22-4 Idiom Explanation

Question:

Explain the meaning of the following idiom: (Idiom)

Answer:
Metaphor for being firm and unshakable.

T23 Riddle

Question:

Guess the riddle: Fireflies, sparkling brightly (Guess a city in China)

Answer:
Kunming(Because "Kunming" literally means "bright insect" in Chinese)

T24 Xiehouyu

Question:
Complete the following Xiehouyu:
—— (The first half of the Xiehouyu)

Answer:
(The second half of the Xiehouyu)

T25 Historical Chinese Phonology

Question:
Which of the following is not a classification name of ancient initials?
A B C D (Options)

Answer: D

T26 Knowledge of Sinology Q&A

Question:

What is the earliest detailed and well-recorded annals history book in ancient times?

Answer:
Zuo Zhuan

Figure 6: Task 22-2 to Task 26 examples.

Table 2: Data source information in WenMind. For details on M1-M5, please refer to Appendix B.3. Zoom in for better view.

ID	Task Name	Data Collection Methods					Data Source	Related Link	License
		M1	M2	M3	M4	M5			
T1-1	Å å í • Inverted Sentence Structure	4					Internet	-	-
T1-2	e å Elliptical Sentence	4					Internet	-	-
T1-3	Å å { ‹ Inverted Sentence Types	4					Internet	-	-
T1-4	$ - å Sentence Structure Identification	4					Internet	-	-
T2	‡ } û Ñ Classical Chinese to Modern Chinese			4			C2MChn [3]	https://github.com/Zongyuan-Jiang/C2MChn	CC BY-NC-ND-4.0
T3	} ‡ û Ñ Modern Chinese to Classical Chinese			4			C2MChn [3]	https://github.com/Zongyuan-Jiang/C2MChn	CC BY-NC-ND-4.0
T4	} ž S Æ + Named Entity Recognition				4		WYWEB [4]	https://github.com/baudzhou/WYWEB	-
T5	å û Punctuation			4			Daizhige [5]	https://github.com/garychowcmu/daizhigev20	-
T6	; ~ { Topic Classification			4			Daizhige [5]	https://github.com/garychowcmu/daizhigev20	-
T7	W Í ã Ê Word Explanation	4					Internet	-	-
T8	û ã Reading Comprehension		4				Internet	-	-
T9	Z Í Function Words	4					Internet	-	-
T10	G W Homophones				4		ACLUE [6]	https://github.com/isen-zhang/ACLUE	CC BY-NC-SA-4.0
T11	U W I Polysemy				4		ACLUE [6]	https://github.com/isen-zhang/ACLUE	CC BY-NC-SA-4.0
T12	‡ ‡™\ Ancient Prose Writing					4	Model	-	-
T13-1	O• ~ Appreciation Exam Questions		4				Internet	http://ts300.5156edu.com/	CC0 1.0
T13-2	ê 1O• Free Appreciation	4					Internet	https://www.gushixuexi.com/	CC0 1.0
T14-1	× \ Poetry Writing					4	Model	-	-
T14-2	Í \ Ci Writing					4	Model	-	-
T14-3	ò \ Qu Writing					4	Model	-	-
T15-1	… ¹î T Content Q&A	4					Internet	https://www.sou-yun.cn/	CC0 1.0
T15-2	˜ î \ î T Title and Author Q&A	4					Internet	https://www.sou-yun.cn/	CC0 1.0
T15-3	åØ™ Write the Next Sentence	4					Internet	https://www.sou-yun.cn/	CC0 1.0
T15-4	åØ™ Write the Previous Sentence	4					Internet	https://www.sou-yun.cn/	CC0 1.0
T15-5	ã' Ø™ Comprehension Dictation		4				Internet	-	-
T15-6	$ - S Á Genre Judgment	4					Internet	https://www.sou-yun.cn/	CC0 1.0
T16	ä × Í û Ñ Ancient Poetry Translation	4					Internet	https://www.sou-yun.cn/	CC0 1.0
T17	Å { Sentiment Classification				4		THU-FSPC [7]	https://github.com/THUNLP-AIPoet/Datasets	-
T18	ä × Íñ ‡ û Ñ Ancient Poetry to English	4					Internet	https://www.sou-yun.cn/	CC0 1.0
T19	× º Ë Í Poet Introduction	4					Internet	https://www.sou-yun.cn/	CC0 1.0
T20	a ã• Analysis of Imagery	4					Internet	https://www.sou-yun.cn/	CC0 1.0
T21-1	¥ T Couplet Following	4					Internet	-	-
T21-2	ùT \ Couplet Writing	4					Internet	-	-
T21-3	ß * y HengPi Writing	4					Internet	-	-
T22-1	Ñ I Í Synonyms	4					Internet	https://www.zdic.net/	CC0 1.0
T22-2	í ú The Origin of Idiom	4					Internet	https://www.zdic.net/	CC0 1.0
T22-3	í t + Idiom Finding	4					Internet	https://www.zdic.net/	CC0 1.0
T22-4	í ã Ê Idiom Explanation	4					Internet	https://www.zdic.net/	CC0 1.0
T23	í Riddle		4				Internet	-	-
T24	G í Xiehouyu	4					Internet	-	-
T25	ãl íóõ Historical Chinese Phonology				4		ACLUE [6]	https://github.com/isen-zhang/ACLUE	CC BY-NC-SA-4.0
T26	ý f 8 Æî T Knowledge of Sinology Q&A	4	4				Internet	-	-

B.3 The Specific Source of the Data

Table 2 presents the data construction methods, data sources, links to the sources, and the licenses of the original data for 42 fine-grained tasks. Data without a specified source fall into three categories: (a) Questions and answers pairs manually constructed using knowledge content from the Internet; (b) Publicly available examination questions; (c) Questions and answers pairs generated by models.

In the data construction method column of Table 2, the following methods are described:

M1: Scraping and creating Q&A pairs from Internet sources; M2: Directly collecting relevant questions from various Internet platforms; M3: Manually constructing Q&A pairs based on existing CCLLA corpora; M4: Utilizing evaluation dataset questions from the CCLLA domain; M5: Using LLM-generated answers for assessment data.

For tasks without specified links, the sources fall into two categories: (a) Data generated by LLMs, providing responses based on training and inference capabilities. (b) Original data derived from widely accessible online knowledge texts, such as high school Chinese language resources, which are not restricted to specific webpages.

B.4 The Construction of Dataset

B.4.1 Specific Process

The process of constructing the dataset is divided into two stages: the data collection process and the data review process.

(1) Data Collection Process

Data collected from the Internet. Collectors: Volunteer A, Volunteer B, and Graduate Student C. Collection process: Based on common tasks in Chinese language exams (reading comprehension, comprehension dictation, *etc*), collect high-quality exam questions made by experts and scholars from the Internet. We require that the questions should be closely matched with the tasks.

Other open-source datasets. Collector: Graduate Student D. Collection process: The data collector conducts research and collection of existing open-source datasets, selects classic tasks and high-quality data in the CCLLA field, and assesses the scarcity and construction difficulty of task data. Ultimately, valuable and currently scarce data are selected for reuse to supplement and improve the evaluation benchmark. Post-processing: (a) Text filtering: Perform operations such as handling of abnormal symbols (*e.g.*, blank squares), deletion of irrelevant content, and after manual inspection, obtain complete and high-quality data. (b) Question writing: Construct Q&A pairs oriented by the needs of different tasks. (c) Standardization of format: Questions for the same task are unified into the same questioning method. We directly set up corresponding question templates according to different tasks.

LLMs. Collector: Graduate Student E. Collection process: Tasks such as ancient poetry and prose writing are inherently open-ended and do not have fixed correct answers. For these tasks, we initially design a range of questions and then employ the ERNIE-3.5 [8] model to generate reference answers. These answers are further reviewed, filtered, and refined through a manual process.

(2) Data Review Process

a. Reviewers: The review is conducted by three individuals, namely Volunteer A, Graduate Student D, and Graduate Student E.

b. A total of approximately 7,000 data entries are collected. The reviewers manually verify the Q&A pairs from three dimensions: the standardization and accuracy of the questions, the correctness and comprehensiveness of the answers, and whether the text content contains ethical issues or unsafe content.

c. Data with non-standard Q&A are deleted or revised, and entries with safety issues are excluded.

d. Finally, the remaining data are rebalanced in terms of task and quality, followed by a second round of proofreading. In this second round, in addition to reviewing the safety and ethical aspects of the data, it is essential to ensure the accuracy and consistency of the Q&A content, with particular attention to grammar, punctuation, and adherence to proper expression standards. The data for the

26 tasks are also rebalanced to prevent excessive or insufficient data for certain tasks. For tasks with less data, additional backup data are incorporated to increase their volume, while for tasks with excessive data, low-quality entries are streamlined to maintain consistent overall quality. Each task's data undergoes at least two rounds of proofreading to ensure relevance, a balanced distribution of task difficulty, and ultimately, the comprehensiveness, accuracy, and safety of the dataset.

B.4.2 Review Details

During the dataset construction process, two rounds of rigorous manual proofreading are conducted to ensure the accuracy and safety of all data. First, the initially collected 7,000 entries undergo the first round of review, where data containing non-standard Q&A pairs, inaccurate or incomplete answers, and entries with potential ethical issues or unsafe content are identified and filtered out. At this stage, 25.20% of the data are manually removed, leaving 5,236 entries. In the second round, the review standards are further refined, focusing not only on identifying additional safety and ethical concerns but also on ensuring the accuracy and consistency of the Q&A pairs, with special attention to the standardization of grammar, punctuation, and expression. An additional 6.89% of the data are removed during this phase, resulting in 4,875 entries that are comprehensive and secure. During this process, supplementary adjustments are made to address any data imbalances across individual tasks.

It is important to note that, despite our strict data review and filtering process, we cannot guarantee the complete absence of errors or safety issues in the dataset. Users should exercise caution and refrain from misusing the data. We emphasize the cultural value of this dataset and hope it provides valuable support and reference for research in the CCLLA field.

B.4.3 Task Collection Requirements and Question Template Explanation

The following presents a simplified version of the task collection requirements and question template explanation adhered to by the collectors during the data collection process for 26 different tasks.

- **T1-Sentence Structure**

  Collection Requirements: Questions are constructed based on specific subcategories (inverted sentence structure, elliptical sentence, inverted sentence types, sentence structure identification). The questions should focus on identifying or analyzing particular features of the sentence structure, such as rearranging the order of inverted sentences, identifying omitted elements, or determining the sentence type. The questions must be clear and explicitly indicate the sentence features or components being examined.

  Template Explanation: The questions are in MCQ format. A classical Chinese sentence is provided, followed by a question that requires selecting the correct answer based on the features of the sentence.

- **T2-Classical Chinese to Modern Chinese**

  Collection Requirements: Select a sentence of classical Chinese and require its translation into modern Chinese. The question must clearly specify the classical Chinese text to be translated. The selected text should have a certain level of difficulty to assess translation skills and comprehension of the classical language.

  Template Explanation: The questions are in QA format. First, the translation task is presented, followed by the classical Chinese sentence that needs to be translated.

- **T3-Modern Chinese to Classical Chinese**

  Collection Requirements: Select sentences in modern Chinese and require their translation into classical Chinese. The question must clearly specify the modern Chinese text to be translated to assess the ability to convert it into classical Chinese.

  Template Explanation: The questions are in QA format. First, the translation task requirements are presented, followed by the modern Chinese text that needs to be converted.

- **T4-Named Entity Recognition**

  Collection Requirements: A sentence of classical Chinese is provided, and the task is to identify named entities within it. The entities include personal names, place names, book titles, or other proper nouns.

Template Explanation: The questions are in QA format. First, the entity recognition task is specified, followed by the classical Chinese text to be analyzed. The answer section lists all identified named entities.

- **T5-Punctuation**

  Collection Requirements: A sentence of classical Chinese that lacks punctuation is selected, and the task requires the subject to add appropriate punctuation marks. The question must clearly specify the text to be punctuated in order to assess understanding of the structure and content of the classical language.

  Template Explanation: The questions are in QA format. The prompt presents a punctuation-free classical Chinese sentence and requires the addition of the corresponding punctuation marks. The answer section provides the complete text with punctuation added.

- **T6-Topic Classification**

  Collection Requirements: A sentence of classical Chinese is provided, and the task requires classifying its content into the provided topic options. The question must include a clear list of topic classification options to assess the subject's understanding of the text's theme.

  Template Explanation: The questions are in QA format. The prompt includes a classical Chinese sentence without a theme label, and the subject is required to select the most relevant topic type from the given multiple-choice options. The answer section should contain the correct topic classification.

- **T7-Word Explanation**

  Collection Requirements: A word or phrase that requires explanation is selected from a classical Chinese sentence, and the task requires the subject to provide its modern Chinese explanation. The question must include the complete classical Chinese sentence to provide context for the word meaning explanation.

  Template Explanation: The questions are in QA format. First, a request for the explanation of the word or phrase is presented, followed by the classical Chinese sentence containing the word or phrase to be explained. Finally, the answer should provide the meaning of the selected word or phrase.

- **T8-Reading Comprehension**

  Collection Requirements: A passage of classical Chinese is selected, and relevant questions are designed to require the subject to respond based on the content of the text. The questions should focus on comprehension skills, such as extracting key information, analyzing character actions, or interpreting event outcomes.

  Template Explanation: The questions are in QA format. The prompt includes a passage of classical Chinese text along with a question related to its content, and the answer should provide a clear explanation or conclusion.

- **T9-Function Words**

  Collection Requirements: A sentence of classical Chinese containing key function words is selected, and the task requires the subject to determine the specific usage of the function word. The question must provide multiple options covering different usages of the function word to test understanding of its specific meaning in the context of classical Chinese.

  Template Explanation: The questions are in MCQ format. A classical Chinese sentence containing the target function word is presented, and the subject is required to choose the most appropriate usage from several explanation options. The answer section should include the correct option.

- **T10-Homophones**

  Collection Requirements: A sentence of classical Chinese containing homophones is selected, and the task requires the subject to clearly identify which character in the provided options is a homophone. The question must present multiple options, each containing a character that needs to be identified as a homophone.

  Template Explanation: The questions are in MCQ format. Several classical Chinese sentences containing homophones are provided, and the subject is required to choose the homophone from the options. The character to be identified is highlighted in the options, and the answer section includes the correct option.

- **T11-Polysemy**

Collection Requirements: A Chinese character with multiple meanings is selected, and questions are constructed based on the various interpretations of this character in different sentences. The questions may require selecting the explanation that corresponds to a given meaning from a specific sentence or identifying the usage that expresses a particular meaning among multiple sentences. This type of question assesses students' ability to understand the context of polysemous words.

Template Explanation: The questions are in MCQ format. The first format presents a specific sentence containing the polysemous character and requires the subject to choose the option that best matches a specific explanation of the character in that sentence. The options list the different meanings of the polysemous character. The second format provides multiple sentences with the polysemous character highlighted and requires the subject to identify the sentence that corresponds to the meaning described in the question. Each option demonstrates the usage of the polysemous character in different contexts.

- **T12-Ancient Prose Writing**

  Collection Requirements: Tasks should be designed to involve writing in classical Chinese across various scenarios or themes, requiring the subject to compose in classical Chinese. These scenarios can include storytelling, writing social media copy, or marketing content. The aim is to assess the subject's ability to create prose in classical Chinese and their skill in using ancient language to express modern ideas flexibly.

  Template Explanation: The questions are in QA format. The prompt is typically presented in the form of a first-person request or command, asking for the creation of a specific type of content, such as a story or copywriting, with the requirement to use classical Chinese. The question structure includes a detailed scenario setup and specific writing instructions, and the answer usually provides a complete prose text in classical Chinese.

- **T13-Appreciation**

  Collection Requirements: Questions should focus on the interpretation and analysis of ancient Chinese poetry. The questions may ask participants to identify incorrect interpretations or analyze imagery, metaphors, and the emotions conveyed in the poem. Through these tasks, subjects need to demonstrate a deep understanding of the language, structure, and cultural background of the poetry.

  Template Explanation: The questions are in QA and MCQ formats. In the first format (MCQ), the question provides a classical poem and annotations, with the options presenting different interpretations or analyses. The subject must select the best answer. In the second format (QA), the question provides a poem along with an open-ended question, asking the subject to analyze the metaphors, imagery, and emotions expressed in the poem. The answer should include an analysis from one or more perspectives, explaining the deeper meanings and emotions of the poem.

- **T14-Ancient Poetry Writing**

  Collection Requirements: Tasks should involve the creation of ancient Chinese poetry or lyrical works. The question must provide a clear theme and specify the required form, such as the genre of the poem, the name of the tune, or the title of the melody. Participants are required to compose within the traditional metrical structure, assessing their language expression skills and mastery of ancient poetry forms.

  Template Explanation: The questions are in QA format. The prompt usually takes the form of a direct request, requiring the creation of a poem, lyric, or song based on a specified theme and format. The question description includes the theme and specific creative form, while the answer should provide a complete example that meets the task requirements.

- **T15-Basic Q&A**

  Collection Requirements: Tasks should cover fundamental questions on various aspects of ancient Chinese poetry. These may include questions on full-text content, identifying the poet and title, completing the next or previous line, understanding-based completion, and determining the genre. The questions must clearly specify the requirements, demonstrating the participant's knowledge of specific aspects of classical poetry.

  Template Explanation: (a) Content Q&A: The questions are in QA format. The question directly asks for the full content of a given poem and requires participants to write it out completely. (b) Title and Author Q&A: The questions are in QA format. The task presents a part of a classical poem and asks the participant to identify its title and author. (c) Write the Next Sentence: The questions are in QA format. The question provides the first line and asks the participant to write the next line of the poem. (d) Write the Previous Sentence: The questions are in QA format. The question gives

the second line and asks the participant to fill in the previous one. (e) Comprehension Dictation: The questions are in FB format. The task provides a hint, such as the poem's theme or function, and asks for a key line from the poem to be filled in. (f) Genre Judgment: The questions are in QA format. The question presents a poem and requires the participant to determine its genre.

- **T16-Ancient Poetry Translation**

  Collection Requirements: The task should involve translating ancient Chinese poetry into modern vernacular Chinese. The question aims to assess the understanding of ancient poetic language and the ability to accurately convert it into more accessible modern language, ensuring the translation conveys the original imagery and emotions.

  Template Explanation: The questions are in QA format. The question asks subject to translate the given classical poem into vernacular Chinese. It includes the classical poem text, and the answer provides the modern vernacular translation of the poem.

- **T17-Sentiment Classification**

  Collection Requirements: The task should assess the ability to understand and classify the sentiment of ancient Chinese poetry. Each question provides a poem and asks the subject to choose the most appropriate sentiment classification from the given options. These options typically include negative, slightly negative, neutral, slightly positive, and positive.

  Template Explanation: The questions are in MCQ format. The question provides the title, author, and content of the poem, then requires the subject to select the sentiment classification that best matches the emotional tone. The options range from "negative" to "positive", and the answer identifies the best choice.

- **T18-Ancient poetry to English**

  Collection Requirements: The task is designed to test the ability to translate ancient Chinese poetry into English. Each question should provide a specific ancient poem and require the subject to translate it into modern English, ensuring that the translation preserves the original imagery and emotional expression of the poem.

  Template Explanation: The questions are in QA format. The task presents an ancient poem and asks for its translation into English. The question section includes the given poem, while the answer section provides a complete English translation.

- **T19-Poet introduction**

  Collection Requirements: The task is designed to guide the introduction of the biographies, achievements, and influences of ancient poets. Each question should present a clear request for detailed information about the specified poet, including their background, representative works, and literary contributions.

  Template Explanation: The questions are in QA format. The question asks for the background and achievements of a particular poet. The question section consists of a simple request for an introduction, with the specific poet designated by the creator. The answer provides a detailed biography of the poet, including birth and death years, major experiences, literary achievements, and their historical and cultural significance.

- **T20-Analysis of imagery**

  Collection Requirements: The task is designed to analyze the multiple meanings of common imagery found in ancient poetry or prose. The questions should highlight specific imagery and require an explanation of its traditional and symbolic significance in different historical texts, thereby assessing the depth of understanding of the imagery.

  Template Explanation: The questions are in QA format. The question asks for a list of the multiple meanings typically associated with the imagery. The question section includes the name of the imagery, while the answer section provides a detailed explanation of the imagery within the context of ancient texts, potentially listing various meanings along with historical background and classic citation examples.

- **T21-Couplet**

  Collection Requirements: The task is designed to test or inspire the understanding and skills in couplet following, couplet writing, or HengPi writing. The questions may involve providing the lower line for a given upper line, composing a couplet based on a festival or theme, or creating a suitable HengPi for an existing couplet. The task assesses mastery of parallel structure and creative interpretation of the couplet's theme.

Template Explanation: The questions are in QA format. The task may ask for couplet following by matching a lower line to an upper line, couplet writing based on a specific festival or theme, or HengPi writing for a given couplet. The question provides the necessary information and challenge, while the answer presents a couplet, matching lower line, or HengPi that meets the specified requirements.

- **T22-Idiom**

  Collection Requirements: The task is designed to assess understanding, explanation, origin, and association of idioms. These questions may involve identifying synonyms of idioms, providing an idiom explanation, finding the origin of the idiom, or locating idioms embedded in a text and explaining their meaning. This task demonstrates the subject's familiarity with idioms and their ability to apply them.

  Template Explanation: The questions are in QA format. The question format involves identifying synonyms of a given idiom, providing an idiom explanation, stating the origin of idiom, or finding idioms in a text and explaining them. Each question provides an idiom or text, and the answer includes relevant idiom information such as synonyms, explanations, origin, or the idioms embedded in the text and their meanings.

- **T23-Riddle**

  Collection Requirements: The task is designed to present engaging and challenging riddles, covering a variety of subjects or themes. The goal is to test the subject's associative thinking, breadth of knowledge, and problem-solving skills. Each riddle should provide sufficient clues for logical reasoning to arrive at the correct answer.

  Template Explanation: The questions are in QA format. The subject is typically asked to guess an ancient historical figure, place name, country, or other objects. The question consists of the riddle and specific requirements, while the answer section provides the correct solution.

- **T24-Xiehouyu**

  Collection Requirements: The task is designed to assess the understanding and recall of Xiehouyu. The question usually provides the first part of the Xiehouyu, and the subject is required to complete the second part. This type of question evaluates familiarity with traditional Chinese cultural expressions of humor and wisdom, as well as the ability to interpret metaphors.

  Template Explanation: The questions are in QA format. The question is presented as an incomplete Xiehouyu, usually giving only the first part, and the subject is asked to complete the saying. The question clearly specifies the need for completion, and the answer section provides the second part of the Xiehouyu, optionally with an explanation.

- **T25-Historical Chinese phonology**

  Collection Requirements: The task involves designing multiple-choice questions on the topic of historical Chinese phonology. These questions require the subject to have a basic understanding of ancient Chinese phonetic structures, including initials and finals, as well as the Qieyun system. The subject must be able to accurately recognize and evaluate phonological phenomena and classical works in ancient Chinese phonology.

  Template Explanation: The questions are in MCQ format. Each question presents a specific knowledge point related to ancient Chinese phonology, along with several options, requiring the subject to select the correct answer. Each question includes a prompt and several answer choices, and the answer section identifies the correct option.

- **T26-Knowledge of Sinology Q&A**

  Collection Requirements: The task involves designing questions covering a wide range of common knowledge in Sinology, including history, geography, linguistics, philosophy, and other aspects of ancient Chinese culture. The questions should be concise and clearly formulated to assess the subject's understanding of traditional Chinese culture.

  Template Explanation: The questions are in QA and MCQ formats. Each question should clearly present a specific topic related to Sinology in a straightforward manner. The question section contains a clear prompt, while the answer section provides a concise and accurate response.

C Evaluation

C.1 Evaluated Model

Details of all evaluated models are shown in Table 3.

Table 3: Details of all evaluated models. Zoom in for better view.

Model	#Parameter	Base Model	Creator	Access	Website	Domain
Baichuan2-7B-Chat [9]	7B	-	Baichuan	Weights	https://huggingface.co/baichuan-inc/Baichuan2-7B-Chat	General
Baichuan2-13B-Chat [9]	13B	-	Baichuan	Weights	https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat	General
Fire y-Baichuan2-13B [10]	13B	Baichuan2-13B-Base [9]	YeungNLP	Weights	https://huggingface.co/YeungNLP/ re y-baichuan2-13b	General
InternLM2-Chat-7B [11]	7B	-	Internlm	Weights	https://huggingface.co/internlm/internlm2-chat-7b	General
Qwen1.5-0.5B-Chat [12]	0.5B	-	Alibaba	Weights	https://huggingface.co/Qwen/Qwen1.5-0.5B-Chat	General
Qwen1.5-4B-Chat [12]	4B	-	Alibaba	Weights	https://huggingface.co/Qwen/Qwen1.5-4B-Chat	General
Qwen1.5-7B-Chat [12]	7B	-	Alibaba	Weights	https://huggingface.co/Qwen/Qwen1.5-7B-Chat	General
Qwen1.5-14B-Chat [12]	14B	-	Alibaba	Weights	https://huggingface.co/Qwen/Qwen1.5-14B-Chat	General
Qwen1.5-32B-Chat [12]	32B	-	Alibaba	Weights	https://huggingface.co/Qwen/Qwen1.5-32B-Chat	General
Qwen1.5-72B-Chat [12]	72B	-	Alibaba	Weights	https://huggingface.co/Qwen/Qwen1.5-72B-Chat	General
Yi-1.5-6B-Chat [13]	6B	-	01.AI	Weights	https://huggingface.co/01-ai/Yi-1.5-6B-Chat	General
Yi-1.5-9B-Chat [13]	9B	-	01.AI	Weights	https://huggingface.co/01-ai/Yi-1.5-9B-Chat	General
Yi-1.5-34B-Chat [13]	34B	-	01.AI	Weights	https://huggingface.co/01-ai/Yi-1.5-34B-Chat	General
ChatGLM2-6B [14]	6B	-	Tsinghua	Weights	https://huggingface.co/THUDM/chatglm2-6b	General
ChatGLM3-6B [14]	6B	-	Tsinghua	Weights	https://huggingface.co/THUDM/chatglm3-6b	General
Ziya-LLaMA-13B-v1.1 [15]	13B	Llama-13B [16]	IDEA	Weights	https://huggingface.co/IDEA-CCNL/Ziya-LLaMA-13B-v1.1	General
LLaMA2-7B-Chat [17]	7B	-	Meta	Weights	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf	General
LLaMA2-13B-Chat [17]	13B	-	Meta	Weights	https://huggingface.co/meta-llama/Llama-2-13b-chat-hf	General
LLaMA2-Chinese-7B-Chat [18]	7B	LLaMA2-7B [17]	FlagAlpha	Weights	https://huggingface.co/FlagAlpha/Llama2-Chinese-7b-Chat	General
LLaMA2-Chinese-13B-Chat [19]	13B	LLaMA2-13B [17]	FlagAlpha	Weights	https://huggingface.co/FlagAlpha/Llama2-Chinese-13b-Chat	General
LLaMA3-8B-Instruct [20]	8B	-	Meta	Weights	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct	General
LLaMA3-Chinese-8B-Chat [21]	8B	LLaMA3-8B [20]	Shenzhi-wang	Weights	https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat	General
Gemma-1.1-7B-IT [22]	7B	-	Google	Weights	https://huggingface.co/google/gemma-1.1-7b-it	General
Bloom-7B-Chunhua [23]	7B	Bloom-7B1 [24]	-	Weights	https://huggingface.co/wptoux/bloom-7b-chunhua	Ancient Chinese
Ancient-Chat-LLM-7B [25]	7B	Internlm-7B [11]	-	Weights	https://modelscope.cn/models/HinGwenWoong/ancient-chat-7b	Ancient Chinese
Xunzi-Qwen1.5-7B [26]	7B	Qwen-1.5-7B [12]	Nanjing Agricultural University	Weights	https://modelscope.cn/models/Xunzillm4cc/Xunzi-Qwen1.5-7B_chat	Ancient Chinese
ERNIE-3.5-8K-0329 [8]	-	-	Baidu	API	https://console.bce.baidu.com/qianfan	General
ERNIE-4.0-8K-0329 [8]	-	-	Baidu	API	https://console.bce.baidu.com/qianfan	General
Spark-3.5 [27]	-	-	iFLYTEK	API	https://xinghuo.xfyun.cn/sparkapi	General
GPT-3.5 [28]	-	-	OpenAI	API	https://chat.openai.com/	General
GPT-4 [29]	-	-	OpenAI	API	https://chat.openai.com/	General

C.2 Scoring Prompt

Scoring prompts of various tasks are illustrated in Figure 7.

C.3 Scoring Consistency Analysis

When evaluating models using scoring methods, it is common to conduct an analysis of the consistency between model scores and human assessments [30, 31, 32, 33]. For the model scoring results on the WenMind benchmark, we have chosen to use the “Agreement Rate” metric to determine whether the model scores align with human expectations.

Agreement Rate. We present the question, reference answer, model output, model score, and the reasoning behind the model's score to humans. The humans then decide whether they agree based on the “score” and the “reasoning”. If the model's score is reasonable and aligns with human understanding and expectations, they indicate “Agreement”. If there are scoring errors or contradictory reasons, they indicate “Disagreement”. The overall consistency between the model's scores and human expectations is determined by the proportion of “Agreements”. This method, compared to having humans score or rank directly for comparison, more directly re ects consistency, reduces the in uence of human subjectivity, speeds up the annotation process, and minimizes errors due to the dif culty of human ranking or scoring.

We randomly sample 417 data points (strati ed by tasks) from the scoring results of ve representative LLMs, and three volunteers perform “Consistency Judgments” on these samples. We then average the “Agreement Rate” of all volunteers. According to Table 4, the average agreement rate across the ve LLMs is 89.4%. This indicates that our model scoring method has a high level of consistency with human expectations, providing results that are of signi cant reference value. In the end, we opts for ERNIE-3.5 [8] as the scoring model, with the cost of completing one round of scoring being approximately $3.8, striking a balance between cost and effectiveness.

Figure 7: Scoring prompt samples.

Table 4: The rate of human agreement with the model scoring.

Model	Overall	Domain			Capability		
		Ancient Prose	Ancient Poetry	Ancient Literary Culture	Understanding	Generation	Knowledge
Baichuan2-7B-Chat [9]	0.890	0.905	0.886	0.873	0.920	0.700	0.938
GPT-3.5 [28]	0.909	0.912	0.913	0.900	0.909	0.857	0.931
LLaMA3-Chinese-8B-Chat [21]	0.882	0.891	0.894	0.855	0.893	0.743	0.931
Qwen1.5-7B-Chat [12]	0.882	0.912	0.881	0.845	0.914	0.686	0.931
Spark-3.5 [27]	0.906	0.878	0.881	0.982	0.898	0.786	0.969
Average	0.894	0.899	0.891	0.891	0.907	0.754	0.940

Table 5: Traditional metrics on translation and punctuation tasks. See Appendix B.1 for details on the tasks represented by T2, T3, T5, T16, and T18. Zoom in for better view.

Model	T2				T3				T16				T18				T5
	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	F1
Baichuan2-7B-Chat [9]	56.5	28.9	53.4	17.8	53.4	23.9	51.3	11.7	48.3	17.5	40.0	9.2	78.3	57.5	73.7	36.9	62.0
Baichuan2-13B-Chat [9]	53.0	24.3	49.8	13.3	49.2	19.6	47.5	7.8	47.1	16.5	38.4	8.8	78.5	57.9	73.8	38.2	59.8
Fire y-Baichuan2-13B [10]	54.2	28.7	51.6	17.5	35.8	16.2	34.3	5.9	52.0	21.5	44.9	11.6	79.2	57.6	73.5	36.8	49.7
ChatGLM2-6B [14]	44.6	19.3	40.9	8.8	41.2	14.7	39.1	4.6	42.6	11.7	32.5	4.1	65.1	48.5	59.6	27.0	37.2
ChatGLM3-6B [14]	52.6	24.5	49.0	13.1	53.3	23.5	51.0	9.6	47.1	15.4	37.6	7.1	72.3	53.6	67.9	31.6	61.9
InternLM2-Chat-7B [11]	60.3	33.0	57.6	21.1	64.2	36.4	61.9	22.1	49.5	17.6	41.1	8.8	77.1	56.7	72.2	35.7	63.0
Qwen1.5-05B-Chat [12]	56.3	27.5	53.4	14.5	54.4	24.6	52.5	11.1	50.6	19.7	46.9	8.8	69.7	47.6	65.1	25.8	30.8
Qwen1.5-4B-Chat [12]	60.0	32.1	57.2	19.7	62.3	33.9	60.0	19.4	53.9	21.3	46.4	11.8	75.8	54.5	71.3	33.9	59.6
Qwen1.5-7B-Chat [12]	54.5	27.5	50.8	25.4	56.6	27.8	53.8	29.1	46.8	15.3	37.4	10.7	73.8	55.1	68.3	27.0	55.4
Qwen1.5-14B-Chat [12]	55.4	28.0	52.0	15.9	57.7	29.2	55.0	17.0	45.5	14.6	35.7	6.6	72.6	54.0	67.9	33.7	61.9
Qwen1.5-32B-Chat [12]	61.0	33.1	57.8	21.4	63.7	36.6	61.3	22.7	50.7	18.3	42.2	10.1	75.7	55.4	70.5	35.3	64.9
Qwen1.5-72B-Chat [12]	58.2	30.2	54.6	17.9	61.2	32.8	58.5	19.4	47.1	15.9	37.4	7.5	78.0	57.8	73.5	37.3	61.9
Yi-1.5-6B-Chat [13]	51.4	24.7	47.4	13.7	50.6	21.5	48.0	8.6	49.4	17.2	40.3	8.4	76.1	55.0	70.8	32.9	50.2
Yi-1.5-9B-Chat [13]	53.4	26.7	50.1	15.7	57.7	28.3	55.2	13.8	48.1	17.1	39.3	8.4	75.9	53.7	69.8	32.1	60.1
Yi-1.5-34B-Chat [13]	54.1	28.3	50.6	16.1	55.9	27.4	53.5	13.1	46.6	16.0	37.3	7.0	78.0	57.3	72.3	35.8	60.7
ERNIE-3.5-8K-0329 [8]	53.0	26.1	49.1	15.2	54.3	30.0	51.5	19.0	36.9	9.1	26.2	1.4	76.5	55.8	70.5	35.6	72.4
ERNIE-4.0-8K-0329 [8]	58.7	31.3	55.6	19.8	60.1	38.8	57.7	28.0	45.6	13.7	35.5	5.6	70.1	52.9	65.6	34.1	61.2
Spark-3.5 [27]	61.5	35.2	58.4	24.4	58.8	35.7	56.2	25.1	48.3	16.3	39.3	7.6	49.3	35.5	45.3	21.0	76.2
Gemma-11-7b-IT [22]	32.4	10.3	28.7	2.0	32.8	10.3	30.6	1.9	44.7	12.9	35.7	4.2	75.3	54.3	69.9	30.0	13.4
Ziya-LLaMA-13B-v1.1 [15]	63.1	35.5	60.4	23.4	39.4	18.6	38.2	9.0	53.3	20.9	46.5	11.4	77.4	55.7	73.3	34.0	57.4
LLaMA2-7B-Chat [17]	10.1	1.3	9.8	0.0	15.0	2.5	14.3	0.0	3.2	0.4	3.0	0.0	72.4	51.5	66.7	13.2	0.1
LLaMA2-13B-Chat [17]	10.4	1.3	10.1	0.0	14.9	2.7	14.3	0.0	17.9	2.2	16.8	0.0	65.3	49.8	59.9	20.0	1.6
LLaMA2-Chinese-7B-Chat [18]	19.6	6.9	17.6	2.0	44.5	19.1	42.9	6.0	10.3	3.0	9.2	0.6	74.2	52.1	67.5	27.8	20.2
LLaMA2-Chinese-13B-Chat [19]	43.8	16.8	40.0	7.8	52.2	23.0	49.9	10.0	44.1	13.3	34.6	5.1	72.1	51.3	67.4	30.9	51.4
LLaMA3-8B-Instruct [20]	15.3	3.0	14.5	0.3	28.9	6.3	27.4	0.0	18.9	3.6	17.0	0.2	60.0	49.4	56.2	14.3	11.1
LLaMA3-Chinese-8B-Chat [21]	50.9	23.2	47.5	12.2	49.5	20.5	47.6	8.4	47.8	15.4	38.6	7.3	71.9	55.7	67.3	35.2	57.7
GPT-3.5 [28]	51.2	22.8	47.5	12.3	48.1	18.0	45.4	7.2	48.7	16.6	40.1	8.5	78.6	57.7	73.9	38.2	58.0
GPT-4 [29]	46.9	20.4	43.2	9.6	50.6	21.1	48.6	9.5	42.8	12.2	32.5	4.0	78.4	56.8	71.8	35.8	62.3
Ancient-Chat-LLM-7B [25]	58.8	31.2	56.3	19.3	61.3	34.8	59.4	22.0	48.8	16.5	41.8	7.6	73.4	46.8	68.8	22.3	67.3
Bloom-7B-Chunhua [23]	55.4	26.4	52.4	14.5	56.0	27.3	53.6	11.8	49.7	18.1	42.4	8.8	75.1	54.7	71.3	32.8	62.0
Xunzi-Qwen1.5-7B [26]	65.9	38.5	63.5	25.4	68.6	43.1	66.6	29.1	49.3	20.5	44.9	10.7	67.7	47.3	63.8	27.0	75.6

C.4 Model Response

We present example responses from Spark-3.5 [27], GPT-4 [29], Qwen1.5-72B-Chat [12], LLaMA3-Chinese-8B-Chat [21], and Xunzi-Qwen1.5-7B [26] models on a subset of tasks, as illustrated in Figure 9, 10, 11, 12 and 13.

C.5 Traditional Metrics

Traditional metrics for all evaluated models on the translation and punctuation tasks are listed in Table 5.

C.6 The Analysis of Problem Dif culty and Performance

When constructing the benchmark, we have additionally annotated “dif culty” information for tasks within the eld of Ancient Poetry, including Content Q&A, Title and Author Q&A, Write the Next Sentence, and Write the Previous Sentence. The dif culty levels are categorized into “Easy” and “Dif cult”. We have used the “Chinese Primary and Secondary School Curriculum” as the basis for classi cation, designating poems and verses covered in the curriculum as “Easy” and those that are uncommon, rare, and outside of the curriculum as “Dif cult”.

The performance of 31 LLMs on easy and dif cult questions is illustrated in Table 6. We also present the variance in performance across different levels of question dif culty using box plots (Figure 8). The highest performance of LLMs on easy questions can reach up to 93.0, whereas the average score on dif cult questions is only 2.1. From this, we can draw a preliminary conclusion that the performance of LLMs on knowledge-type tasks is highly dependent on the scope of the pre-training data. Most LLMs are trained with data that includes some common and simple ancient poetry content, but there is a severe lack of comprehensive and extensive ancient poetry data. It is worth mentioning that the Spark-3.5 model [27] achieved a score of 17.2 on dif cult questions, signi cantly outperforming other models. As a general model, it has considerable potential in the eld of ancient poetry.

Table 6: Performance of LLMs on easy and dif cult problems. Zoom in for better view.

Model	Overall		Content Q&A		Title and Author Q&A		Write the Next Sentence		Write the Previous Sentence	
	Simple	Hard	Simple	Hard	Simple	Hard	Simple	Hard	Simple	Hard
Baichuan2-7B-Chat [9]	52.0	0.0	75.6	0.0	40.0	0.0	62.5	0.0	29.2	0.0
Baichuan2-13B-Chat [9]	61.5	0.3	92.8	0.0	48.3	0.4	79.2	0.0	20.8	1.3
Fire y-Baichuan2-13B [10]	71.2	0.1	97.7	0.2	52.5	0.0	91.7	0.0	50.0	0.0
ChatGLM2-6B [14]	38.7	0.4	54.5	0.1	25.0	0.4	62.5	0.0	20.8	1.3
ChatGLM3-6B [14]	36.7	0.6	50.9	0.0	19.2	0.4	66.7	0.0	25.0	2.6
InternLM2-Chat-7B [11]	79.1	4.2	83.7	0.0	75.7	1.3	91.7	13.2	66.7	9.2
Qwen1.5-0.5B-Chat [12]	2.9	0.0	4.5	0.0	4.2	0.0	0.0	0.0	0.0	0.0
Qwen1.5-4B-Chat [12]	67.2	0.5	97.7	0.0	65.8	0.0	66.7	0.0	16.7	2.6
Qwen1.5-7B-Chat [12]	79.1	1.4	100.0	0.1	67.5	0.8	83.3	1.3	66.7	5.3
Qwen1.5-14B-Chat [12]	80.3	5.3	99.4	0.0	82.5	0.0	79.2	13.2	41.7	18.4
Qwen1.5-32B-Chat [12]	87.4	3.0	98.8	0.8	75.8	0.4	100.0	3.9	83.3	11.2
Qwen1.5-72B-Chat [12]	85.4	4.5	100.0	0.0	73.3	0.0	95.8	10.5	79.2	15.8
Yi-1.5-6B-Chat [13]	67.3	1.9	78.1	0.0	58.3	0.3	87.5	3.9	50.0	6.6
Yi-1.5-9B-Chat [13]	74.9	3.8	82.7	0.0	69.2	2.9	87.5	9.2	62.5	7.9
Yi-1.5-34B-Chat [13]	87.9	2.7	92.8	0.1	86.4	2.1	95.8	7.9	75.0	3.9
ERNIE-3.5-8K-0329 [8]	93.0	5.8	99.7	0.6	86.7	2.8	100.0	9.2	89.6	18.4
ERNIE-4.0-8K-0329 [8]	91.1	5.1	100.0	0.4	82.5	1.0	100.0	10.5	87.5	17.1
Spark-3.5 [27]	83.7	17.2	99.7	46.1	77.5	1.2	91.7	3.9	62.5	0.0
Gemma-1.1-7B-IT [22]	2.6	0.0	0.0	0.0	6.7	0.0	0.0	0.0	0.0	0.0
Ziya-LLaMA-13B-v1.1 [15]	11.8	0.1	19.5	0.2	9.2	0.0	16.7	0.0	0.0	0.0
LLaMA2-7B-Chat [17]	0.7	0.3	0.0	0.0	1.7	1.0	0.0	0.0	0.0	0.0
LLaMA2-13B-Chat [17]	0.3	0.3	0.0	0.1	0.8	0.2	0.0	1.3	0.0	0.0
LLaMA2-Chinese-7B-Chat [18]	0.0	0.1	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0
LLaMA2-Chinese-13B-Chat [19]	0.3	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0
LLaMA3-8B-Instruct [20]	1.7	0.4	0.3	0.0	2.5	0.0	4.2	0.0	0.0	1.3
LLaMA3-Chinese-8B-Chat [21]	2.4	1.7	2.7	0.0	0.8	3.9	8.3	1.3	0.0	1.3
GPT-3.5 [28]	20.1	0.0	41.5	0.0	0.8	0.0	45.8	0.0	4.2	0.0
GPT-4 [29]	55.7	3.0	91.1	0.1	36.7	3.0	75.0	6.6	20.8	5.3
Ancient-Chat-LLM-7B [25]	54.5	0.7	88.6	0.0	26.9	2.4	75.0	0.0	41.7	0.0
Bloom-7B-Chunhua [23]	30.9	0.0	71.4	0.0	5.0	0.0	45.8	0.0	8.3	0.0
Xunzi-Qwen1.5-7B [26]	72.6	0.6	92.2	1.3	71.7	0.4	79.2	0.0	33.3	0.0
Average	48.2	2.1	61.8	1.6	40.5	0.8	57.8	3.1	33.4	4.2

Figure 8: Box plot of LLMs' scores on easy and dif cult questions. “Min” represents the minimum value; “Max” represents the maximum value; “Q3” represents the upper quartile; “Q1” represents the lower quartile.

Figure 9: The sample responses of LLMs on the sentence structure task.

Figure 10: The sample responses of LLMs on the basic Q&A task.

Figure 11: The sample responses of LLMs on the knowledge of Sinology Q&A task.

Figure 12: The sample responses of LLMs on the xiehouyu task.

Figure 13: The sample responses of LLMs on the classical Chinese to modern Chinese task.

C.7 The Analysis of Fine-Grained Tasks and Performance

Due to space limitations, the main body of the paper only presents the metrics of LLMs on 26 coarse-grained tasks. In fact, we have subdivided six of these tasks (T1-Sentence structure, T13-Appreciation, T14-Ancient poetry writing, T15-Basic Q&A, T21-Couplet, T22-Idiom) into subtasks, resulting in a total of 42 ne-grained tasks. Table 7 provides the performance of LLMs on these subtasks (ne-grained). Different subtasks involve more detailed knowledge points and methods of examination, hence there are signi cant performance differences among the subtasks. Overall, LLMs generally perform poorly on tasks involving elliptical sentence, content Q&A of ancient poetry, title and author Q&A of ancient poetry, write the previous sentence, and the origin of idiom. Researchers can use it as a reference to focus on improving the capabilities of LLMs in certain areas.

The task numbers presented in Table 7 correspond to the following meanings, with detailed information available in Appendix B.1.

T1-1: Inverted Sentence Structure; T1-2: Elliptical Sentence; T1-3: Inverted Sentence Types; T1-4: Sentence Structure Identi cation; T13-1: Appreciation Exam Questions; T13-2: Free Appreciation; T14-1: Poetry Writing; T14-2: Ci Writing; T14-3: Qu Writing; T15-1: Content Q&A; T15-2: Title and Author Q&A; T15-3: Write the Next Sentence; T15-4: Write the Previous Sentence; T15-5: Comprehension Dictation; T15-6: Genre Judgment; T21-1: Couplet Following; T21-2: Couplet Writing; T21-3: HengPi Writing; T22-1: Synonyms; T22-2: The Origin of Idiom; T22-3: Idiom Finding; T22-4: Idiom Explanation.

Table 7: Performance of LLMs on ne-grained tasks. Zoom in for better view.

Model	Ancient Prose				Ancient Poetry											Ancient Literary Culture						
	T1-1	T1-2	T1-3	T1-4	T13-1	T13-2	T14-1	T14-2	T14-3	T15-1	T15-2	T15-3	T15-4	T15-5	T15-6	T21-1	T21-2	T21-3	T22-1	T22-2	T22-3	T22-4
Baichuan2-7B-Chat [9]	20.4	10.9	0.0	60.5	50.4	60.7	52.3	59.0	69.0	16.2	12.0	15.0	7.0	49.4	31.7	52.0	59.6	69.8	36.7	2.5	24.6	52.3
Baichuan2-13B-Chat [9]	36.1	15.6	42.9	79.1	55.2	60.1	59.8	65.3	64.2	19.9	14.8	19.0	6.0	46.7	35.0	51.0	61.5	55.4	19.8	2.5	36.7	60.3
Fire y-Baichuan2-13B [10]	17.1	10.9	42.9	11.6	39.5	49.2	34.3	44.0	40.5	21.1	15.8	22.0	12.0	63.3	18.3	29.2	68.7	69.2	31.6	5.8	27.2	40.3
ChatGLM2-6B [14]	8.3	7.8	28.6	69.8	33.5	54.8	50.3	56.0	60.5	11.8	7.8	15.0	6.0	50.0	14.2	48.5	48.7	57.6	29.0	1.0	23.5	39.9
ChatGLM3-6B [14]	25.0	7.8	57.1	76.7	52.3	59.8	49.7	61.2	55.0	10.9	6.0	16.0	8.0	38.3	8.3	42.8	65.5	60.4	21.8	1.0	24.7	47.5
InternLM2-Chat-7B [11]	38.9	20.3	14.3	67.4	64.1	74.2	43.7	61.6	53.5	18.0	23.2	32.0	23.0	57.2	50.8	60.6	75.6	70.0	50.5	7.0	61.4	72.3
Qwen1.5-0.5B-Chat [12]	24.1	3.1	28.6	11.4	30.5	23.0	20.0	15.4	13.0	1.0	1.3	0.0	0.0	11.7	12.5	47.3	61.2	62.7	2.3	5.0	12.2	17.3
Qwen1.5-4B-Chat [12]	35.2	17.2	42.9	51.2	46.1	46.8	29.7	45.8	33.0	21.0	19.8	16.0	6.0	66.7	31.7	47.8	70.4	61.6	10.9	8.0	32.2	45.1
Qwen1.5-7B-Chat [12]	38.0	17.2	28.6	74.4	67.4	70.2	53.3	69.4	74.5	16.6	20.8	21.0	20.0	68.3	28.3	49.0	74.3	67.8	41.8	1.5	31.5	72.8
Qwen1.5-14B-Chat [12]	47.2	23.4	0.0	46.5	72.8	73.0	55.7	69.5	70.7	21.4	24.8	29.0	24.0	83.3	50.8	49.9	76.8	68.2	45.0	2.5	44.8	70.9
Qwen1.5-32B-Chat [12]	45.4	42.2	28.6	27.9	77.2	73.5	55.0	66.9	73.5	21.9	23.0	27.0	28.5	90.0	56.7	50.8	78.2	67.4	65.1	5.3	53.9	67.8
Qwen1.5-72B-Chat [12]	69.4	34.4	85.7	39.5	79.7	74.1	57.3	69.6	76.5	21.5	22.0	31.0	31.0	89.2	58.8	52.7	79.4	68.5	53.1	10.0	45.3	78.2
Yi-1.5-6B-Chat [13]	25.0	14.1	0.0	65.1	62.6	68.0	53.0	64.5	65.5	16.8	17.7	24.0	17.0	70.0	15.0	46.1	72.2	66.1	49.8	3.5	45.8	63.1
Yi-1.5-9B-Chat [13]	46.8	20.3	42.9	44.2	62.4	72.6	58.0	63.6	68.7	17.8	22.8	28.0	21.0	69.4	32.5	49.8	77.0	69.9	69.2	7.0	51.3	67.3
Yi-1.5-34B-Chat [13]	69.4	23.4	71.4	69.8	70.2	73.5	61.0	68.5	70.2	20.0	27.4	29.0	21.0	73.3	51.7	69.2	72.6	70.7	82.6	7.2	61.0	73.3
ERNIE-3.5-8K-0329 [8]	72.2	37.5	85.7	58.1	76.0	75.4	59.0	69.2	63.5	21.9	28.0	31.0	35.5	93.3	67.5	72.0	84.3	83.8	86.9	24.0	57.2	77.9
ERNIE-4.0-8K-0329 [8]	83.3	43.8	85.7	34.9	79.7	78.9	62.2	67.2	67.0	21.8	25.4	32.0	34.0	100.0	68.3	78.8	81.7	83.6	87.3	22.5	58.8	79.7
Spark-3.5 [27]	58.3	20.3	57.1	27.9	68.8	64.5	56.3	59.7	72.5	17.6	24.1	25.0	15.0	86.7	44.2	59.0	69.5	73.5	88.6	75.2	59.2	98.0
Gemma-1.1-7B-IT [22]	27.8	4.7	0.0	9.3	24.1	38.2	32.3	50.4	53.0	0.0	2.0	0.0	0.0	5.0	2.9	37.1	49.8	61.4	5.0	5.0	16.3	8.3
Ziya-LLaMA-13B-v1.1 [15]	30.6	0.0	71.4	60.5	43.9	49.1	39.0	53.8	57.0	4.3	2.8	4.0	0.0	26.7	18.3	44.5	53.2	52.8	13.2	5.0	29.8	48.8
LLaMA2-7B-Chat [17]	0.0	0.0	0.0	0.0	24.1	27.7	16.3	45.6	46.5	0.0	1.2	0.0	0.0	0.0	0.0	4.0	42.3	11.5	4.0	0.0	15.0	16.0
LLaMA2-13B-Chat [17]	10.2	3.1	0.0	9.3	27.6	33.2	26.0	49.8	67.0	1.0	4.0	10.0	0.0	5.6	1.7	26.5	53.3	21.8	5.2	5.0	18.0	27.8
LLaMA2-Chinese-7B-Chat [18]	2.8	0.0	0.0	7.0	16.1	27.2	16.7	25.6	5.0	0.0	3.0	0.0	0.0	0.0	0.0	10.0	30.9	31.0	4.5	0.0	8.3	10.9
LLaMA2-Chinese-13B-Chat [19]	18.5	1.6	14.3	0.0	24.3	32.1	19.7	24.4	20.2	0.0	3.0	0.0	0.0	6.7	2.1	30.7	30.1	57.7	3.6	0.0	5.5	17.2
LLaMA3-8B-Instruct [20]	15.3	1.6	0.0	69.8	44.0	51.6	31.0	57.2	68.2	1.0	8.0	20.0	10.0	3.3	6.7	46.1	64.2	67.7	24.2	5.0	22.8	41.5
LLaMA3-Chinese-8B-Chat [21]	22.2	0.0	0.0	69.8	47.7	55.8	46.3	54.4	51.0	6.0	3.0	3.0	10.0	1.7	8.0	52.7	49.7	66.6	24.7	5.0	25.0	37.1
GPT-3.5 [28]	22.2	15.6	14.3	72.1	43.4	52.0	44.3	57.8	64.0	8.9	3.0	11.0	10.0	6.7	10.8	38.2	51.7	52.6	5.2	1.5	14.8	44.8
GPT-4 [29]	52.8	18.8	57.1	16.3	60.8	69.4	62.7	64.7	68.7	19.7	13.1	23.0	9.0	13.3	19.7	45.4	73.9	56.4	49.5	4.0	36.3	64.8
Ancient-Chat-LLM-7B [25]	25.0	6.3	14.3	76.7	24.9	11.0	30.3	35.4	21.5	19.0	9.8	18.0	10.0	43.3	10.0	45.3	42.3	53.1	4.1	5.0	18.3	43.1
Bloom-7B-Chunhua [23]	13.9	3.1	57.1	23.3	23.4	35.6	45.7	37.8	33.0	15.4	1.5	11.0	2.0	23.9	13.3	41.3	48.8	52.4	17.1	3.5	25.7	42.8
Xunzi-Qwen1.5-7B [26]	26.9	15.6	0.0	65.1	39.0	18.2	46.0	39.6	36.0	20.8	21.8	19.0	8.0	76.7	1.7	52.7	44.7	55.9	23.9	1.0	40.0	30.8
Average	33.2	14.2	31.3	47.3	49.4	53.3	44.1	54.0	54.3	14.6	12.7	16.3	11.2	45.8	24.7	46.2	61.7	60.2	34.1	6.4	33.1	50.3

C.8 Performance Radar Chart

Performance radar charts of all evaluated LLMs are depicted from Figure 14 through 44. The radar charts from left to right represent the elds of Ancient Prose, Ancient Poetry, and Ancient Literary Culture. The meanings of T1-T26 are detailed in Appendix B.1.

C.9 Analysis of Model Scoring Preferences

We design two methods to analyze the preference bias in model scoring.

Use another model to scoreGPT-3.5-Turbo-0125[28] is used as an alternative scoring model for the ve evaluated models. The results are shown in Table 8. The results show that the ranking of models remain the same as when scored by ERNIE-3.5-8K-0329 [8], with ERNIE-4.0-8K-0329 [8] still ranking rst. This suggests that its leading position is genuine and less affected by scoring bias.

Manual analysis. We categorize the questions in WenMind into objective and subjective types, and conduct manual analysis on the model's scoring results for each. We nd that for objective questions with standard answers, ERNIE essentially judges and scores each one strictly according to the scoring points of the answer, showing a weaker bias. For subjective questions, we select classical Chinese writing and Qu writing tasks for analysis. (1) In classical Chinese writing task, ERNIE-4.0-8K-0329

scores 70.8, ranking second among all models. The rst place is the Spark-3.5[27] model with a score of 71.2. We review the response of ERNIE-4.0-8K-0329 and nd that ERNIE-4.0-8K-0329 indeed has strong capabilities in classical Chinese writing, but this does not rule out the in uence of model bias. (2) In the Qu writing task, ERNIE-4.0-8K-0329 scores 67, ranking 11th among all models. This also indirectly indicates that in subjective tasks, ERNIE does not necessarily give high scores to its own answers.

In summary, the preference bias in model scoring is relatively weak, making the use of model scoring a viable alternative to manual scoring and traditional metrics.

Table 8: The results of scoring ve models using ERNIE-3.5-8K-0329 and GPT-3.5-Turbo-0125, respectively.

	ERNIE-3.5-8K-0329	GPT-3.5-Turbo-0125
ERNIE-4.0-8K-0329[8]	64.3	79.7
Spark-3.5[27]	60.9	77.4
Qwen1.5-14B-Chat[12]	54.9	73.7
Baichuan2-13B-chat[9]	45.5	66.4
Xunzi-Qwen1.5-7B[26]	37.0	59.8

C.10 Error Analysis of LLMs

We observe signi cant performance differences in the model across different tasks. We conduct an error analysis on T1 (Sentence Structure), T13 (Appreciation), T14 (Ancient Poetry Writing), T15 (Basic Q&A), T21 (Couplet), and T22 (Idiom), which involve more specialized sub-tasks.

T1-Sentence Structure.The model performs weakest on Elliptical Sentence (T1-2), with a score of only 14.2. It struggles with identifying the omitted parts and restoring sentence completeness. In contrast, its performance on Sentence Structure Identi cation (T1-4) is better, with a score of 47.3, indicating a relative strength in recognizing simpler sentence structures. However, the model's overall understanding and application of various sentence types remain incomplete, especially when dealing with the implicit semantics and grammatical features of classical Chinese.

T13-Appreciation. The model achieves a higher score in Free Appreciation (T13-2) compared to Appreciation Exam Questions (T13-1), with scores of 53.3 and 49.4, respectively. This suggests that the model performs better in less restrictive contexts, showing adaptability in textual analysis. In structured exam questions, however, its limitations become apparent, likely due to the deeper cultural understanding and appreciation skills required, which the model has not yet fully mastered.

T14-Ancient Poetry Writing. The model scores well in Ci Writing (T14-2) and Qu Writing (T14-3), with scores of 54.0 and 54.3, indicating strong performance in writing tasks with xed formats. However, its score for Poetry Writing (T14-1) is lower, at 44.1, re ecting the challenge of producing creative and artistically expressive poetry, particularly when dealing with diverse themes and styles.

T15-Basic Q&A. The lowest score is observed in Title and Author Q&A (T15-2), with only 12.7, highlighting challenges in recalling and recognizing speci c textual information, particularly in large datasets. In contrast, Comprehension Dictation (T15-5) achieves the highest score, at 45.8, indicating the model's relative strength in language comprehension and summarization, despite its struggles with more speci c questions.

T21-Couplet. The model excels in Couplet Writing (T21-2) and HengPi Writing (T21-3), scoring 61.7 and 60.2, respectively, demonstrating its capability in generating structured and formatted content. However, it performs less well in Couplet Following (T21-1), with a score of 46.2, indicating dif culties in completing couplets, where maintaining contextual harmony and creativity is crucial.

T22-Idiom. The lowest score is in The Origin of Idiom (T22-2), with only 6.4, revealing the model's signi cant shortcomings in associating idioms with their origins, re ecting a lack of historical knowledge. By contrast, it performs better in Idiom Explanation (T22-4), with a score of 50.3, showing that the model can provide accurate explanations for common idioms, though deeper cultural and contextual understanding still needs improvement.

In summary, the model performs well in tasks with more rigid rules and structures but struggles in tasks requiring a strong cultural background and high creativity, highlighting areas for future improvement.

C.11 Traditional Chinese Version of WenMind

We additionally provide a traditional Chinese version of the WenMind dataset to evaluate the performance of LLMs in the context of traditional Chinese within the CCLLA eld. Table 9, 10, and 11 show the evaluation results for the traditional Chinese version of WenMind. The Spark-3.5 model [27] had been discontinued at the time of the traditional Chinese evaluation experiments, so no corresponding evaluation results are available.

Below is a brief analysis of the evaluation results:

Overall Performance and Trends. The average score for the traditional Chinese version is 36.7, which is lower than the 41.2 for the simpli ed Chinese version. This gap may re ect limitations in the model's performance when processing traditional Chinese texts, potentially in uenced by the language environment and modern usage frequency. The main tasks focus on classical texts and ancient poetry, indicating the advantage of traditional Chinese in traditional cultural domains. However, its performance is somewhat lacking in modern applications and data support.

Model Performance. In the traditional Chinese version, ERNIE-4.0-8K-0329 [8] leads with a total score of 58.4, demonstrating strong performance in handling tasks related to traditional Chinese. In the simpli ed Chinese version, this model also excels, with a total score of 64.3, maintaining a top position. This re ects ERNIE-4.0-8K-0329's consistently high performance across both versions. In contrast, LLaMA2-7B-Chat [17] scores only 14.2 in the traditional Chinese evaluation, making it the weakest model. In the simpli ed Chinese version, this model also scores poorly, with a score of 13.0. This indicates LLaMA2-7B-Chat's consistent underperformance in both language versions, likely due to insuf cient targeted training data to handle the complexity and diversity of the language.

Parameter Scale and Model Ef ciency. As the model parameter size increases, there is a noticeable improvement in performance. For example, the performance of the Qwen1.5 series [12] improves with an increase in parameters, particularly in complex text tasks. This suggests that large-scale models are better at handling traditional Chinese tasks by offering enhanced data processing and feature extraction capabilities, providing performance improvements across various tasks. This conclusion is consistent with the ndings for the simpli ed Chinese version.

Table 9: Results of all evaluated models on the traditional Chinese version of WenMind. Zoom in for better view.

Model	Overall	Ancient Prose												Ancient Poetry										Ancient Literary Culture						
		T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25	T26
Baichuan2-7B-Chat [9]	30.0	17.7	48.1	53.3	53.6	38.5	23.8	23.2	49.6	35.0	31.5	35.5	34.7	20.3	42.5	2.8	49.4	54.0	63.3	53.0	10.4	43.3	18.5	8.0	8.0	24.0	41.5			
Baichuan2-13B-Chat [9]	43.2	48.5	59.3	59.9	66.6	66.5	42.0	39.5	61.4	40.0	23.5	50.5	57.7	45.6	49.0	16.1	54.4	49.0	50.2	19.0	45.6	57.8	45.3	13.0	18.5	38.0	59.9			
Fire y-Baichuan2-13B [10]	38.5	23.7	61.0	62.2	65.9	74.0	34.0	35.2	57.2	31.0	23.5	49.5	56.5	32.0	35.4	15.0	59.6	30.5	51.9	13.2	18.6	55.0	41.2	5.0	9.0	32.0	58.6			
ChatGLM2-6B [14]	33.0	35.3	47.9	56.4	63.7	50.2	30.0	37.0	53.0	23.0	4.5	34.0	58.0	34.3	41.1	10.1	47.1	23.0	36.1	12.2	29.4	57.8	34.0	9.0	4.5	20.0	44.1			
ChatGLM3-6B [14]	35.1	43.2	47.0	58.6	49.6	58.2	31.0	41.3	52.1	33.0	26.0	36.0	60.9	36.4	41.0	8.4	45.2	36.4	43.3	13.2	30.1	60.8	37.0	7.0	4.0	31.0	36.5			
InternLM2-Chat-7B [11]	43.9	28.2	56.0	56.9	67.3	73.8	45.5	49.5	70.2	43.0	26.5	49.0	62.6	45.0	46.9	16.5	56.7	48.3	55.8	26.0	49.4	61.4	47.3	8.0	10.5	26.0	58.5			
Qwen1.5-0.5B-Chat [12]	30.9	48.5	49.7	56.0	65.4	57.2	36.0	18.7	42.2	28.0	19.0	42.0	50.6	21.5	42.6	4.5	55.7	35.5	26.2	1.6	3.8	55.6	29.2	0.0	7.0	22.0	15.4			
Qwen1.5-4B-Chat [12]	36.9	34.0	49.8	57.2	54.6	69.9	31.0	25.2	50.5	48.0	19.0	52.0	55.3	33.7	36.6	16.0	50.3	46.3	43.2	10.7	20.4	60.3	31.2	2.0	8.5	27.0	54.2			
Qwen1.5-7B-Chat [12]	42.8	40.0	51.0	55.6	53.4	68.8	40.0	59.2	76.5	61.0	13.0	55.5	65.1	54.1	54.6	15.5	52.7	59.0	51.6	26.1	46.8	55.9	35.1	13.0	9.0	33.0	57.5			
Qwen1.5-14B-Chat [12]	49.3	31.3	57.3	57.8	54.8	73.6	45.0	63.8	75.3	69.0	29.0	61.0	68.8	56.8	23.4	58.8	62.0	59.1	31.2	62.3	62.9	49.7	11.0	14.5	33.0	72.1				
Qwen1.5-32B-Chat [12]	50.7	30.3	62.3	62.6	54.6	77.0	42.5	59.0	71.5	64.0	30.5	68.5	69.5	61.3	54.5	24.8	57.9	58.7	57.0	38.1	65.5	61.1	48.6	23.5	16.0	46.0	76.5			
Qwen1.5-72B-Chat [12]	53.4	40.5	60.2	61.1	55.9	79.4	50.2	72.0	79.8	67.0	39.5	71.0	69.8	64.2	56.9	25.4	63.3	64.5	58.1	36.5	64.6	62.1	53.1	24.0	13.0	43.0	84.1			
Yi-1.5-6B-Chat [13]	43.0	30.3	55.7	61.5	61.3	66.2	42.0	61.8	69.8	53.0	19.0	49.0	66.6	45.2	51.1	11.9	57.2	43.8	51.5	30.0	54.4	63.4	46.3	7.0	2.5	26.0	62.3			
Yi-1.5-9B-Chat [13]	43.9	29.5	60.0	58.6	49.4	68.3	44.0	70.0	73.8	58.0	24.0	58.0	64.5	46.1	51.3	12.8	58.4	51.9	54.9	27.6	47.3	60.8	45.4	8.0	6.5	30.0	63.0			
Yi-1.5-34B-Chat [13]	51.8	35.0	66.1	60.3	54.5	72.9	53.0	68.0	79.8	76.0	31.0	62.0	66.4	55.4	52.1	24.6	63.6	59.7	58.2	38.3	73.3	65.9	49.8	15.0	8.0	47.0	79.5			
ERNIE-3.5-8K-0329 [8]	56.7	53.2	62.7	55.1	64.8	81.4	43.3	72.0	79.3	64.0	36.5	72.5	64.0	58.1	49.9	33.4	63.4	59.4	57.5	34.8	65.7	75.1	53.8	68.0	43.0	57.0	85.2			
ERNIE-4.0-8K-0329 [8]	58.4	40.2	62.8	52.9	62.4	82.4	53.8	70.5	75.3	74.0	48.0	74.5	64.2	61.0	50.5	32.9	59.0	65.9	58.0	25.1	65.5	70.8	55.3	84.0	67.5	72.0	88.3			
Gemma-1.1-7B-IT [22]	23.3	8.1	24.4	42.1	41.9	29.5	26.0	16.8	27.9	18.0	17.0	35.5	55.8	19.6	36.1	0.7	24.3	46.0	38.2	1.1	16.1	52.3	27.1	2.0	0.0	20.0	5.6			
Ziya-LLaMA-13B-v1.1 [15]	31.4	35.6	45.2	61.7	16.8	56.0	27.5	33.8	53.9	31.0	26.5	31.0	54.8	23.8	33.5	4.5	49.0	42.5	40.4	11.2	28.1	58.8	37.9	3.0	4.5	16.0	34.0			
LLaMA2-7B-Chat [17]	14.2	3.0	1.9	2.8	15.9	0.6	14.5	2.0	23.3	10.0	1.0	8.0	53.9	22.1	29.7	0.2	17.4	32.2	29.6	0.0	6.3	65.2	23.0	0.0	0.0	5.0	2.5			
LLaMA2-13B-Chat [17]	20.2	5.0	14.2	23.1	30.2	19.1	12.6	14.8	40.9	20.0	3.0	15.0	57.3	22.1	32.9	1.4	36.0	38.5	40.6	3.0	22.1	53.7	29.5	2.0	0.0	10.0	5.4			
LLaMA2-Chinese-7B-Chat [18]	19.5	4.7	13.7	38.5	51.8	19.8	17.2	13.5	28.2	20.0	2.5	23.5	56.1	15.2	36.0	0.2	21.3	18.4	30.5	2.8	11.3	53.3	28.8	0.0	0.0	19.0	3.1			
LLaMA2-Chinese-13B-Chat [19]	21.1	9.7	27.7	39.6	42.8	14.3	14.5	15.7	34.9	21.0	12.5	25.0	52.0	19.1	31.6	0.7	27.0	27.3	38.1	3.1	12.4	51.1	30.3	0.0	0.0	13.0	11.5			
LLaMA3-8B-Instruct [20]	31.6	26.8	41.8	50.2	41.5	48.6	43.5	42.7	56.7	39.0	13.5	44.0	57.4	30.6	30.0	1.8	42.9	45.0	46.4	4.0	35.9	57.3	33.5	4.0	0.0	24.0	22.1			
LLaMA3-Chinese-8B-Chat [21]	35.8	36.3	43.0	60.4	47.4	70.2	46.0	49.5	61.1	41.0	22.5	49.5	61.9	35.8	39.0	1.1	51.9	43.3	47.3	5.7	43.8	55.4	39.7	3.0	2.0	21.0	21.8			
GPT-3.5 [28]	37.7	37.5	47.4	60.5	47.2	71.1	43.0	38.3	55.3	34.0	25.5	49.0	62.2	43.9	43.5	4.5	50.9	50.5	55.9	9.1	37.5	58.4	41.8	6.0	0.0	31.0	36.3			
GPT-4 [29]	46.2	17.0	56.2	62.3	51.4	73.1	47.5	65.8	80.2	52.0	33.0	69.5	67.9	55.6	51.1	11.1	58.0	57.0	56.6	22.8	60.4	61.4	46.5	17.0	7.0	39.0	63.6			
Ancient-Chat-LLM-7B [25]	15.9	8.0	4.5	2.2	52.2	2.4	7.5	5.0	13.3	13.0	21.0	20.0	36.8	9.0	31.9	5.0	18.7	5.5	4.7	2.7	2.6	67.2	25.6	0.0	1.0	9.0	11.0			
Bloom-7B-Chunhua [23]	25.8	28.5	25.4	23.2	51.1	75.1	20.5	27.2	43.3	31.0	8.0	39.5	27.4	21.9	17.6	6.5	43.8	39.5	41.2	11.5	9.8	41.0	16.1	4.0	9.0	26.0	37.9			
Xunzi-Qwen1.5-7B [26]	35.3	40.2	59.4	51.7	26.2	81.5	50.5	22.5	53.3	53.0	18.5	49.0	63.1	36.0	49.5	16.8	57.7	8.0	44.2	14.9	0.7	55.9	32.9	6.0	7.5	37.0	52.7			
Average	36.7	29.0	45.4	50.1	50.5	57.3	34.6	40.5	56.3	41.7	21.6	46.0	58.1	37.6	42.5	11.6	48.4	43.4	46.0	18.0	34.7	58.7	37.8	11.8	9.4	29.2	44.8			

D Data Access

Our benchmark and related code are available at https://github.com/SCUT-DLVCLab/WenMind .

Table 10: Results of all evaluated models on different domains and capabilities in the traditional Chinese version of WenMind.

Model	Overall	Domain			Capability		
		Ancient Prose	Ancient Poetry	Ancient Literary Culture	Understanding	Generation	Knowledge
Baichuan2-7B-Chat [9]	30.0	38.4	23.7	26.4	39.1	41.5	14.8
Baichuan2-13B-Chat [9]	43.2	51.8	33.6	44.4	50.9	56.0	29.6
Fire y-Baichuan2-13B [10]	38.5	49.7	26.2	40.0	46.9	51.4	23.9
ChatGLM2-6B [14]	33.0	41.0	23.2	35.4	38.6	54.5	19.7
ChatGLM3-6B [14]	35.1	44.4	24.4	37.1	42.3	56.8	19.8
InternLM2-Chat-7B [11]	43.9	52.8	34.8	43.7	51.6	58.7	29.5
Qwen1.5-0.5B-Chat [12]	30.9	44.1	18.1	29.5	41.5	52.0	11.1
Qwen1.5-4B-Chat [12]	36.9	46.3	27.4	36.6	44.9	54.6	21.4
Qwen1.5-7B-Chat [12]	42.8	51.4	36.4	38.8	51.8	57.5	26.8
Qwen1.5-14B-Chat [12]	49.3	56.1	43.4	47.8	56.6	62.9	36.1
Qwen1.5-32B-Chat [12]	50.7	57.4	44.4	49.8	57.4	61.5	38.9
Qwen1.5-72B-Chat [12]	53.4	61.2	46.2	52.0	61.5	62.6	40.1
Yi-1.5-6B-Chat [13]	43.0	52.2	33.3	43.5	50.6	61.6	28.0
Yi-1.5-9B-Chat [13]	43.9	53.7	34.1	43.4	52.8	59.6	27.9
Yi-1.5-34B-Chat [13]	51.8	59.2	44.9	50.5	58.9	63.2	39.2
ERNIE-3.5-8K-0329 [8]	56.7	61.3	47.7	63.7	60.8	67.9	48.2
ERNIE-4.0-8K-0329 [8]	58.4	63.0	47.5	68.3	62.6	65.4	50.9
Gemma-1.1-7B-IT [22]	23.3	29.4	15.2	26.1	28.5	49.7	9.3
Ziya-LLaMA-13B-v1.1 [15]	31.4	38.9	21.4	35.0	37.9	52.9	16.9
LLaMA2-7B-Chat [17]	14.2	9.6	11.5	26.2	11.9	55.8	6.1
LLaMA2-13B-Chat [17]	20.2	19.6	16.9	26.4	21.6	50.3	10.2
LLaMA2-Chinese-7B-Chat [18]	19.5	24.0	10.5	26.4	21.3	50.4	8.7
LLaMA2-Chinese-13B-Chat [19]	21.1	25.6	12.9	26.8	24.4	47.4	9.6
LLaMA3-8B-Instruct [20]	31.6	41.5	21.1	32.1	40.3	51.9	14.7
LLaMA3-Chinese-8B-Chat [21]	35.8	48.8	23.7	33.5	46.8	53.4	16.4
GPT-3.5 [28]	37.7	48.2	26.9	37.8	47.7	56.2	19.4
GPT-4 [29]	46.2	56.3	36.2	45.6	55.9	60.6	29.4
Ancient-Chat-LLM-7B [25]	15.9	15.6	8.2	29.1	13.3	54.0	9.2
Bloom-7B-Chunhua [23]	25.8	33.8	18.4	24.4	34.3	33.6	12.4
Xunzi-Qwen1.5-7B [26]	35.3	45.6	23.7	37.0	41.8	56.0	21.0
Average	36.7	44.0	27.9	38.6	43.2	55.3	23.0

Table 11: Traditional metrics on translation and punctuation tasks in the traditional Chinese version of WenMind. See Appendix B.1 for details on the tasks represented by T2, T3, T5, T16, and T18. Zoom in for better view.

Model	T2				T3				T16				T18				T5
	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	F1
Baichuan2-7B-Chat [9]	38.1	15.3	36.4	6.8	41.5	17.3	39.6	5.9	29.7	9.9	26.5	3.8	44.5	23.5	40.1	8.3	18.4
Baichuan2-13B-Chat [9]	49.4	21.4	46.3	9.8	57.7	28.3	55.8	13.5	49.1	19.4	40.4	7.9	78.8	57.2	73.3	37.5	50.7
Fire y-Baichuan2-13B [10]	52.7	24.2	50.9	10.9	53.9	25.5	52.4	11.9	50.6	19.7	47.4	8.6	78.5	56.4	72.9	35.5	50.5
ChatGLM2-6B [14]	44.6	18.4	38.9	6.2	41.8	16.8	40.4	4.4	41.0	12.9	33.3	4.3	69.1	50.8	63.7	28.8	33.9
ChatGLM3-6B [14]	47.8	20.3	43.9	8.1	48.5	20.6	47.0	7.7	44.2	14.9	36.0	6.1	71.8	53.4	66.4	32.7	39.8
InternLM2-7B-Chat [11]	48.8	22.5	45.1	9.4	56.0	28.5	53.9	13.4	47.9	17.8	38.4	6.2	76.3	55.0	70.5	31.7	58.9
Qwen1.5-0.5B-Chat [12]	51.2	22.3	49.1	10.3	51.9	23.8	50.3	9.8	48.7	18.6	45.3	7.9	69.3	47.4	64.4	25.2	11.2
Qwen1.5-4B-Chat [12]	56.4	27.9	53.7	16.6	58.0	28.8	56.3	14.0	51.5	20.1	47.3	8.6	76.5	54.1	72.0	33.0	48.5
Qwen1.5-7B-Chat [12]	45.1	18.1	42.2	8.0	53.0	23.2	49.9	9.9	42.5	13.2	35.0	6.3	75.7	55.5	69.5	33.1	55.4
Qwen1.5-14B-Chat [12]	52.9	24.3	49.7	13.1	54.4	23.9	51.5	11.1	46.6	15.3	37.9	7.2	72.7	53.7	68.2	32.2	61.4
Qwen1.5-32B-Chat [12]	54.4	26.8	51.6	14.3	59.2	30.7	57.1	15.8	45.3	15.7	37.8	7.6	76.0	56.3	71.0	36.5	67.2
Qwen1.5-72B-Chat [12]	51.3	22.9	48.0	11.7	58.0	29.0	55.8	14.4	46.6	15.4	37.7	7.4	77.3	57.7	73.1	36.7	70.4
Yi-1.5-6B-Chat [13]	53.0	26.8	49.0	14.0	50.6	21.7	48.3	8.3	47.8	16.4	38.3	6.2	75.1	53.1	69.8	31.2	45.2
Yi-1.5-9B-Chat [13]	50.8	22.8	47.1	9.9	56.8	27.5	54.5	12.4	48.2	17.2	38.4	6.9	78.1	56.5	71.8	34.1	56.6
Yi-1.5-34B-Chat [13]	50.0	24.5	45.6	10.3	52.3	24.9	50.6	8.9	48.3	18.4	37.6	7.0	77.6	56.8	71.5	35.7	60.4
ERNIE-3.5-8K-0329 [8]	56.1	26.2	53.8	12.2	59.2	31.1	57.4	16.3	43.4	15.1	31.1	3.9	75.7	56.3	69.0	35.8	69.3
ERNIE-4.0-8K-0329 [8]	53.3	24.7	50.8	11.1	62.6	37.1	61.0	23.5	45.0	15.0	35.7	5.7	69.4	52.7	64.4	31.6	63.7
Gemma-11-7b-IT [22]	35.6	12.1	32.5	3.6	39.2	14.2	37.2	2.7	36.0	9.6	30.0	3.0	73.9	53.5	68.8	29.0	18.1
Ziya-LLaMA-13B-v1.1 [15]	40.8	17.3	39.1	7.4	27.5	8.8	26.6	2.5	41.0	13.3	35.9	5.6	77.5	55.5	72.6	33.6	46.5
LLaMA2-7B-Chat [17]	14.9	2.0	14.3	0.0	11.8	1.5	11.4	0.0	20.0	3.6	18.7	0.1	72.9	52.4	66.5	16.9	0.1
LLaMA2-13B-Chat [17]	18.8	2.9	18.1	0.0	22.0	3.6	21.4	0.0	32.1	6.4	30.0	0.2	72.4	52.5	65.8	17.5	0.7
LLaMA2-Chinese-7B-Chat [18]	25.9	8.2	23.1	0.9	30.3	10.6	29.1	1.9	41.6	14.2	37.1	4.6	72.0	52.5	66.8	25.3	10.0
LLaMA2-Chinese-13B-Chat [19]	28.6	9.1	25.5	1.6	25.7	8.9	24.7	1.6	11.4	2.8	10.2	0.5	74.2	53.4	67.5	29.9	9.5
LLaMA3-8B-Instruct [20]	31.5	8.0	28.5	0.4	30.5	8.1	29.2	0.6	39.7	8.3	36.0	0.6	57.2	48.4	53.8	12.1	9.8
LLaMA3-Chinese-8B-Chat [21]	50.1	21.8	46.3	11.1	55.7	25.5	53.7	11.0	46.1	14.5	53.7	6.1	76.1	56.8	71.4	34.3	53.4
GPT-3.5 [28]	50.2	22.2	46.6	11.2	53.8	24.0	51.9	9.6	46.9	16.2	38.4	7.6	78.7	57.4	72.6	37.1	58.9
GPT-4 [29]	48.9	21.3	46.1	10.0	56.1	25.5	53.5	12.2	45.1	14.0	36.3	6.2	77.1	57.2	72.0	36.1	62.9
Ancient-Chat-LLM-7B [25]	25.9	8.2	23.7	0.1	25.4	9.1	23.8	0.0	33.5	8.8	27.7	0.0	54.3	42.4	51.7	0.3	27.0
Bloom-7B-Chunhua [25]	49.4	21.4	47.4	5.6	48.1	22.0	46.9	3.5	51.1	19.6	47.6	7.1	76.4	54.6	72.2	32.9	48.8
Xunzi-Qwen1.5-7B [26]	62.7	34.9	60.7	22.5	65.7	39.8	64.0	25.9	49.7	20.0	46.2	9.5	76.1	53.1	71.8	31.6	69.8

E Related Work on Ancient Language Evaluation Datasets

In the process of exploring ancient Chinese evaluation benchmarks, we discovered a series of benchmark studies targeting other ancient languages. The AGREE Benchmark [34] focuses on evaluating distributional semantic models for Ancient Greek, the LogogramNLP benchmark [35] makes natural language processing analysis of ancient pictographic languages possible for the first time, the Sanskrit Word Segmentation Dataset [36] provides a foundation for syntactic and semantic processing of Sanskrit literature, and The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE) [37] is a 1.5 million-word syntactically annotated corpus of old English prose, based on the Toronto Corpus of old English, encompassing all major old English prose works. Compared to these benchmarks that include a small number of tasks, WenMind includes a rich variety of ancient Chinese tasks, making it possible to comprehensively evaluate the ancient Chinese capabilities of LLMs.

F Statement of Responsibility

The WenMind benchmark is released under the CC-BY-NC-SA-4.0 license and strictly adheres to the agreements of the original data sources. The licenses for the original data sources are detailed in Appendix B.3. We have reviewed the ethical guidelines and ensured that the content of the paper and the benchmark are in compliance with the guidelines.

Figure 14: Performance of Baichuan2-7B-Chat.

Figure 15: Performance of Baichuan2-13B-Chat.

Figure 43: Performance of Bloom-7B-Chunhua.

Figure 16: Performance of Fire y-Baichuan2-13B.

Figure 17: Performance of ChatGLM2-6B.

Figure 18: Performance of ChatGLM3-6B.

Figure 19: Performance of InternLM2-Chat-7B.

Figure 20: Performance of Qwen1.5-0.5B-Chat.

Figure 21: Performance of Qwen1.5-4B-Chat.

Figure 22: Performance of Qwen1.5-7B-Chat.

Figure 23: Performance of Qwen1.5-14B-Chat.

Figure 24: Performance of Qwen1.5-32B-Chat.

Figure 25: Performance of Qwen1.5-72B-Chat.

Figure 26: Performance of Yi-1.5-6B-Chat.

Figure 27: Performance of Yi-1.5-9B-Chat.

Figure 28: Performance of Yi-1.5-34B-Chat.

Figure 29: Performance of ERNIE-3.5-8K-0329.

Figure 30: Performance of ERNIE-4.0-8K-0329.

Figure 31: Performance of Spark-3.5.

Figure 32: Performance of Gemma-1.1-7B-IT.

Figure 33: Performance of Ziya-LLaMA-13B-v1.1.

Figure 34: Performance of LLaMA2-7B-Chat.

Figure 35: Performance of LLaMA2-13B-Chat.

Figure 36: Performance of LLaMA2-Chinese-7B-Chat.

Figure 37: Performance of LLaMA2-Chinese-13B-Chat.

Figure 38: Performance of LLaMA3-8B-Instruct.

Figure 39: Performance of LLaMA3-Chinese-8B-Chat.

Figure 40: Performance of GPT-3.5.

Figure 41: Performance of GPT-4.

Figure 42: Performance of Ancient-Chat-LLM-7B.

Figure 44: Performance of Xunzi-Qwen1.5-7B.

References

[1] Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. Datasets for large language models: A comprehensive survey. arXiv preprint arXiv:2402.18041, 2024.

[2] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol., 15(3), mar 2024.

[3] Zongyuan Jiang, Jiapeng Wang, Jiahuan Cao, Xue Gao, and Lianwen Jin. Towards better translations from classical to modern Chinese: A new dataset and a new method. In CCF International Conference on Natural Language Processing and Chinese Computing, pages 387–399. Springer, 2023.

[4] Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. WYWEB: A NLP evaluation benchmark for classical Chinese. In Findings of the Association for Computa-

tional Linguistics: ACL 2023, pages 3294–3319, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Daizhigev20. https://github.com/garychowcmu/daizhigev20.

[6] Yixuan Zhang and Haonan Li. Can large langauge model comprehend ancient Chinese? A preliminary test on ACLUE. In Proceedings of the Ancient Language Processing Workshop, pages 80–87, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria.

[7] Huimin Chen, Xiaoyuan Yi, Maosong Sun, Cheng Yang, Wenhao Li, and Zhipeng Guo. Sentiment-controllable Chinese poetry generation. In Proceedings of the Twenty-Eighth International Joint Conference on Arti cial Intelligence, Macao, China, 2019.

[8] Baidu. ERNIE. https://yiyan.baidu.com/.

[9] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305, 2023.

[10] YeungNLP. Fire y-Baichuan2-13B. https://huggingface.co/YeungNLP/ re y-baichuan2-13b, 2024. Hugging Face.

[11] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. InternLM2 technical report, 2024.

[12] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023.

[13] 01. AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01.AI, 2024.

[14] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. GLM: General language model pretraining with autoregressive blank in lling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022.

[15] IDEA-CCNL. Ziya-LLaMA-13B-v1.1. https://huggingface.co/IDEA-CCNL/Ziya-LLaMA-13B-v1.1.

[16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and ef cient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[17] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and ne-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[18] FlagAlpha. Llama2-Chinese-7b-Chat. https://huggingface.co/FlagAlpha/Llama2-Chinese-7b-Chat, 2024. Hugging Face.

[19] FlagAlpha. Llama2-Chinese-13b-Chat. https://huggingface.co/FlagAlpha/Llama2-Chinese-13b-Chat, 2024. Hugging Face.

[20] Meta. LLaMA3. https://ai.meta.com/blog/meta-llama-3.

[21] Shenzhi Wang. Llama3-8B-Chinese-Chat. https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat, 2024. Hugging Face.

[22] Google. Gemma-1.1-7B-IT. https://huggingface.co/google/gemma-1.1-7b-it, 2024. Hugging Face.

[23] Wptoux. Bloom-7B-Chunhua. https://huggingface.co/wptoux/bloom-7b-chunhua, 2024. Hugging Face.

[24] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176B-parameter open-access multilingual language model, 2023.

[25] PeterH0323. Ancient-Chat-LLM. https://github.com/PeterH0323/ancient-chat-llm, 2024. GitHub.

[26] Xunzi-LLM of Chinese-classics. XunziALLM. https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM, 2024. GitHub.

[27] I ytek. Spark-v3.5. https://xinghuo.xfyun.cn/.

[28] OpenAI. GPT-3.5. https://openai.com/index/gpt-3-5-turbo- ne-tuning-and-api-updates.

[29] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. GPT-4 technical report, 2024.

[30] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. Advances in Neural Information Processing Systems, 36, 2023.

[31] Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. Evaluating hallucinations in Chinese large language models. arXiv preprint arXiv:2310.03368, 2023.

[32] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[33] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36, 2023.

[34] Silvia Stopponi, Saskia Peels-Matthey, and Malvina Nissim. AGREE: A new benchmark for the evaluation of distributional semantic models of ancient Greek. Digital Scholarship in the Humanities, 39(1):373–392, 2024.

[35] Danlu Chen, Freda Shi, Aditi Agarwal, Jacobo Myerston, and Taylor Berg-Kirkpatrick. LogogramNLP: Comparing visual and textual representations of ancient logographic writing systems for NLP. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14238–14254, 2024.

[36] Amrith Krishna, Pavankumar Satuluri, and Pawan Goyal. A dataset for Sanskrit word segmentation. In Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 105–114, 2017.

[37] Ann Taylor. The York-Toronto-Helsinki parsed corpus of old English prose. Creating and digitizing language corpora: Volume 2: Diachronic Databases, pages 196–227. Springer, 2003.