Chinese Stylistic Competence: Evaluation Method and Datasets of Large Language Model's Performance

Liwei Zhou
Beijing Language and Culture University
Beijing, China
liweiyeahmail@163.com

Gaoqi Rao*
Beijing Language and Culture University
Beijing, China
raogaoqi@blcu.edu.cn

***Abstract*—Stylistic competence is an important pragmatic competence, and adequate stylistic competence is required for large language model (LLM) to land in language life. In this paper, stylistic competence is defined as the ability to use appropriate style for communication in a specific register, and based on this, three tasks of stylistic classification, stylistic generation, and stylistic transformation are designed to evaluate the Chinese stylistic competence of LLMs represented by ChatGPT. It is found that LLMs have their own advantages and limitations in different tasks and styles. GPT-4 demonstrates the most comprehensive and excellent Chinese stylistic competence, ChatGPT3.5 and ERNIE Bot have better performance, ChatGLM-6B and SparkDesk have unstable performance and notable shortcomings, with their overall abilities being somewhat lackluster. In addition, the degree of informality of the texts generated by each model is relatively limited, the literary grace is ordinary, and problems such as consistency errors, normative errors, factual errors, illogicality, insufficient sentence fluency, and obvious traces of machine translation still exist. For LLM, it should be viewed from an instrumental perspective, expanding its stylistic competence with rich and diversified stylistic data resources and technological advances, and at the same time reasonably utilizing and giving full play to its stylistic resources attribute, to make it better serve the language life.**

Keywords—large language model; stylistic competence; language resource

I. INTRODUCTION

Since the end of 2022, large language model (LLM) represented by ChatGPT has garnered widespread attention. Compared with the general language model, the training corpus of LLM is huge and the parameters are dramatically inflated, which makes LLM produce an amazing knowledge emergence capability. In this context, research on the evaluation of LLM is particularly important, both to affect the development of LLM's technology and application, and to deepen human understanding of LLM.

The training corpus of LLM contains a large number of language expressions and scenes, which makes LLM also emerge a certain stylistic competence. Language life is inseparable from style, and stylistic competence is an essential ability for LLM to integrate into human language life. In terms of evaluation, at present, academics mainly focus on the content-level issues such as the factual consistency and accuracy, and pay less attention to the pragmatic competence represented by stylistic competence, which lacks complete and scientific evaluation methods. This study attempts to explore and practice the evaluation method for the Chinese stylistic competence of LLMs represented by ChatGPT, compare and evaluate their respective advantages and limitations, provide inspiration for the improvement direction of LLMs, and discuss the practical application potential of their Chinese stylistic competence in language life from the perspective of resource construction.

*Corresponding author.

II. RELATED WORKS

A. Stylistic Competence

Scholars have discussed the meaning and performance of stylistic competence from different perspectives. According to Li [1], stylistic competence is expressed as a combination of the ability to recognize context type and to select appropriate language materials and expressions. Wu [2] believes that stylistic competence is an aspect of speech activity competence, including the ability to express and apply a certain style. Zhou [3] argues that stylistic competence is usually composed of four aspects: cognitive ability of context type, mastery ability of stylistic markers, comprehension and expression ability of style, and transformation ability of different styles. The investigation of Chinese stylistic competence mainly focuses on language teaching, especially in the field of international Chinese language education, and the stylistic competence of foreign students is usually examined through corpus analysis, error analysis and questionnaire [4]-[7].

B. Evaluation of LLM

Currently, the evaluation methods for LLM mainly include benchmark evaluation [8]-[10], human evaluation [11] and LLM evaluation [12]-[13]. The evaluation tasks mainly include classic natural language processing tasks, such as sentiment analysis, automatic summarization, question-answering, machine translation [14]-[16], as well as performance tests in specific application scenarios, including medical, financial, educational and other fields [17]-[20].

In summary, the existing evaluations pay insufficient attention to the stylistic competence of LLM, which in a way can reflect its ability of language comprehension, communication and expression in complex real-life scenarios. This study will select typical tasks that can cover stylistic competence and large-scale, high-quality stylistic datasets, to comprehensively examine the Chinese stylistic competence of LLMs such as ChatGPT, providing inspiration for their improvement and application, and also providing methodological references for training and testing human stylistic competence.

III. EVALUATION OF LLM'S STYLISTIC COMPETENCE

979-8-3503-3078-6/23/$31.00

On the basis of predecessors, this paper defines stylistic competence as the ability to use appropriate style for communication in a specific register, including stylistic recognition, comprehension, expression and transformation on the basis of context identification. Formality is the most basic and primitive stylistic category and the essential attribute of discourse [21]. Pragmatic competence is the knowledge about the conditions and ways of appropriate use of language, and stylistic competence is an important basic competence for it.

Stylistic competence needs to be externalized into language behavior to be truly reflected. By providing as many expressive scenarios as possible, and starting from the three typical tasks of stylistic classification, stylistic generation and stylistic transformation, it is feasible to mobilize stylistic knowledge and demonstrate all aspects of stylistic competence in a more comprehensive way. For this study, the stylistic classification task primarily examines LLMs' ability to comprehend and recognize style; the stylistic generation task focuses on LLMs' stylistic expression and language application competence based on stylistic understanding and context recognition; and the stylistic transformation task mainly examines LLMs' stylistic transformation and stylistic expression competence. The three types of tasks have their own emphases and cross each other. Therefore, by synthesizing the performance of LLMs in the above three tasks, a more in-depth, comprehensive and effective understanding and evaluation of their Chinese stylistic competence can be achieved. Simultaneously, this study also introduces the human experts indicator to form a reference and comparison with the high level of human Chinese stylistic competence, which not only enables a more intuitive and profound grasp of the stylistic competence of LLMs, but also a beneficial attempt to evaluate the human stylistic competence.

IV. Methods and Results

In this study, API is used to obtain the data, the *model* is gpt-3.5-turbo and gpt-4, and the *temperature* is set to 1. We compare the data with ChatGLM-6B, ERNIE Bot2.0, and SparkDesk1.0.

A. Dataset

Tai et al. [22] constructed a stylistic classification corpus (SCC) of 1.92 million characters based on 8 typical stylistic corpora of official documents, political comments, academic literature, news, novels, prose, microblogs and lyrics, and formed a computationally based classification system of stylistic categories: At the first level, style is divided into formal style and informal style. Under the formal style, it is divided into guiding and informative style, the former includes official document and academic style, while the latter includes political comment and news style. Under the informal style, it is divided into creative, monologue and dialogue style, containing novel and prose style, microblog and lyrics style, conversation and question-answering style respectively. This stylistic classification system is adopted in this study. Huang et al. [23], based on the *Tongyici Cilin*, artificially constructed 4,438 sets of stylistic chain collocation sentences (SCCS) and other parallel stylistic resources from oral to general to written languages. The present study is evaluated on the basis of the above datasets.

B. Stylistic Classification

The number of texts in SCC is large and long. Considering the limitation of the number of words asked, this task randomly selects 125 texts from each of the 8 types of stylistic corpora, and extracts 150-200 word segments from each text, a total of 1,000 samples for testing. Combined with previous research and multiple attempts, the following prompt is used for this task:

请判断下列文本属于以下 8 个语体中的哪一类，注意这个文本只能属于某一类语体。

文本：{text}

语体：公文、学术文献、政论、新闻报道、小说、散文、微博、歌词

(Please judge which of the following 8 styles the following text belongs to, and note that this text can only belong to a certain style.

Text:{text}

Styles: official document, academic literature, political comments, news, novel, prose, microblog, lyrics)

Since this task has adopted the objective question form with the only answer, the answer is the stylistic classification result of human experts, so the human experts score of this task is unified as 1. The recall rate (R) and F1 score (F1) are selected to evaluate the classification results of LLMs, and the data are shown in Table I.

In the classification of each style, GPT-4 and ChatGPT show superior performance, ERNIE Bot also achieves better classification results, while SparkDesk and ChatGLM-6B have weak classification competence. On the whole, the formal and informal style recognition competence of each model remain largely balanced. Furthermore, it can be found that the recall rate of ChatGPT, GPT-4 and ERNIE Bot on the academic literature are all 1, suggesting that their training corpus encompass a considerable amount of academic literature; the recall rate and F1 score of ChatGLM-6B in prose are both 0, which may be due to the apparent dearth of prose in the training corpus; each model has generally achieved a high score in the lyrics, which is speculated to be related to the unique and distinct structural features of the lyrics. Regarding other types of errors, hallucination, logical incorrectness, inconsistency and other problems are concentrated in ChatGLM-6B, ERNIE Bot, and SparkDesk, reflecting the limitations of LLMs in Chinese comprehension and expression.

C. Stylistic Generation

This task evaluates the generation competence of LLMs in 9 styles: official document, political comment, academic literature, news, novel, prose, microblog, lyrics, and dialogue.

For 8 types of styles such as official document and political comment, ChatGPT randomly generates 10 themes for each, which requires covering as many fields as possible. Combined with the actual application requirements and previous studies, the themes undergo verification, and based on this, the role-playing method is adopted to specify the generated themes. Example of prompt is as follows:

TABLE I. STYLISTIC CLASSIFICATION RESULTS OF LLMS

LLM	Index	Formal Style					Informal Style					AVG
		Official Document	Political Comment	Academic Literature	News	AVG	Novel	Prose	Microblog	Lyrics	AVG	
ChatGPT	R	0.82	0.80	1.00	0.72	0.84	0.96	0.68	0.59	0.81	0.76	0.80
	F1	0.81	0.68	0.95	0.75	0.80	0.85	0.75	0.70	0.89	0.80	0.80
GPT-4	R	0.90	0.74	1.00	0.88	0.88	0.97	0.74	0.77	0.89	0.84	0.86
	F1	0.89	0.76	0.89	0.87	0.85	0.87	0.81	0.85	0.94	0.87	0.86
ChatGLM-6B	R	0.35	0.62	0.10	0.30	0.34	0.02	0.00	0.32	0.54	0.22	0.28
	F1	0.36	0.24	0.18	0.29	0.27	0.03	0.00	0.40	0.70	0.28	0.28
ERNIE Bot	R	0.64	0.70	1.00	0.71	0.76	0.70	0.49	0.38	0.86	0.61	0.68
	F1	0.76	0.52	0.79	0.65	0.68	0.80	0.54	0.55	0.91	0.70	0.69
SparkDesk	R	0.94	0.02	0.03	0.49	0.37	0.36	0.56	0.10	0.18	0.30	0.33
	F1	0.42	0.03	0.06	0.51	0.26	0.51	0.41	0.16	0.31	0.35	0.30

假如你是一位记者，请撰写一篇以"某国家/地区的运动员在国际大赛上夺得金牌"为主题的新闻稿。

(If you are a journalist, please write a news release with the theme of "an athlete from a country/region winning a gold medal in an international competition".)

For the dialogue style, 10 real-life dialogue scenarios with stylistic differentiation are designed by ChatGPT with reference to the four elements of people, things, places and attitudes [24]. Example of prompt is as follows:

假如你是一位销售员，正在商场向顾客介绍产品，顾客询问价格和功能。请生成一段你和顾客的对话。

(If you are a salesperson introducing products to customers in a shopping mall, and customers inquire about prices and functions. Please generate a conversation between you and the customer.)

Due to the uncertainty of the output of LLMs, this study asks LLMs 6 times for each prompt, and each LLM generates a total of 540 texts.

This task adopts the method of human evaluation to appraise the generation competence of LLMs. Referring to the view of Lu [25], this study takes appropriateness, normativeness and literary grace (limited to novel, prose and lyrics) as indicators to measure the quality of the generated texts, and the final score is the average value of each indicator. A total of 6 annotators are recruited for this task, all of whom are undergraduates or postgraduates majoring in linguistics, and each person is assigned to one of the six different texts generated by LLMs under the same prompt. In addition, we randomly insert 45 high-quality real texts as human experts indicator into the generated texts at a ratio of 10:1 without informing the annotators. The topics of the real texts basically come from the topics in LLMs' prompts, and the length of the real texts is approximately the same as that of the generated texts. There are five real texts for each of the 9 styles, and each text is jointly rated by 6 annotators. The annotating specifications are as follows:

a) Indicators and Scores: This study utilizes a five-point scale, with annotators scoring appropriateness, normativeness, and literary grace on a scale of 1-5 respectively, and providing written evaluations. Among them, appropriateness focuses on whether the language style used by the communicator is appropriate to the situation. Normativeness is concerned with whether sentences conform to the modern Chinese language norms, including the presence of grammatical errors, typos, variant characters, variant words; the mixing of simplified characters, traditional characters, English and Pinyin; and the inappropriate use of punctuation. Literary grace is expressed in the beauty of colour, sound, decoration, emotion, image and philosophy [26].

b) Invalid outputs: If LLM does not generate text or the generated text is too short to score, all indicators are assigned 0 points.

c) Special cases: If other redundant information appears in the generated text, it will be ignored when scoring and only the text will be scored. If the text is not completed, scored normally.

d) Misconceptions: Factual errors, irrationality, illogicality and lack of fluency do not affect the score.

e) Other evaluation items: Recording other errors or improprieties (as described in point 4), personal reading feelings, etc., except for the 3 evaluation indicators.

Finally, the average value (AVG) and standard deviation (SD) of the scoring results of each indicator are utilized to evaluate the stylistic generation competence of LLMs. The relevant data are presented in Table II and Table III.

The highest averages are obtained for GPT-4 and ChatGLM-6B under informal and formal style, respectively. In each style, the highest score tends to be achieved by the Chinese LLM, but GPT-4 obtains the highest grand average score of 4.37, proving that its overall performance is more stable and superior. In informal style, LLMs still have a certain gap compared with human experts, while in formal style, LLMs' generation competence in official document, academic, and news style has reached a level comparable to or even surpassing that of human experts. Specifically for the various styles, we note that: In the prose style, formulaic expressions such as "firstly", "secondly", "lastly", etc. recur, and the style of the texts tends to be argumentative, which is excessively formal. In the dialogue style, LLMs perform better in scenes with higher formality, such as railway stations and banks, while in scenes with extremely low formality, such as conversations between couples, mother and child, the generated sentences are overly formal. ERNIE Bot and SparkDesk may be affected by sensitive words in prompts, generating a large amount of invalid texts in official document and political comment style. From the three indicators, the average scores of appropriateness and normativeness of each model typically reach 4 points, i.e., t-

TABLE II. HUMAN SCORING RESULTS OF INFORMAL STYLE GENERATED TEXTS IN LLMS

Indicator	Text Source	Novel		Prose		Lyrics		Microblog		Dialogue		AVG	
		AVG	SD	AVG	SD	AVG	SD	AVG	SD	AVG	SD	AVG	SD
Appropriateness	human experts	4.93	0.25	4.73	0.57	4.93	0.25	4.93	0.25	4.80	0.48	4.87	0.36
	ChatGPT	4.50	0.59	2.95	0.97	4.78	0.41	4.52	0.62	3.83	1.02	4.12	0.72
	GPT-4	4.75	0.72	4.45	0.76	4.77	0.50	4.83	0.37	3.77	1.12	4.51	0.69
	ChatGLM-6B	4.45	0.62	2.73	0.87	4.68	0.56	4.42	0.67	3.30	1.01	3.92	0.75
	ERNIE Bot	4.68	0.56	3.42	1.13	4.78	0.45	4.62	0.61	3.50	1.32	4.20	0.82
	SparkDesk	4.67	0.77	2.73	0.96	4.75	0.50	4.67	0.60	3.98	0.89	4.16	0.74
Normativeness	human experts	4.87	0.34	4.53	0.72	4.73	0.44	4.73	0.51	4.77	0.42	4.73	0.49
	ChatGPT	3.52	1.18	4.08	0.82	4.73	0.48	4.52	0.56	4.75	0.47	4.32	0.70
	GPT-4	3.65	1.20	3.90	0.85	4.58	0.64	4.53	0.67	4.90	0.30	4.31	0.73
	ChatGLM-6B	4.75	0.47	4.68	0.53	4.82	0.50	4.67	0.57	4.73	0.51	4.73	0.52
	ERNIE Bot	4.78	0.41	4.77	0.74	4.95	0.22	4.68	0.53	4.62	1.25	4.76	0.63
	SparkDesk	4.57	0.80	4.85	0.40	4.98	0.13	4.82	0.47	4.80	0.40	4.80	0.44
Literary Grace	human experts	4.87	0.43	4.50	0.85	4.17	0.78	-	-	-	-	4.51	0.68
	ChatGPT	3.23	0.80	3.15	0.65	3.65	0.81	-	-	-	-	3.34	0.76
	GPT-4	3.90	1.12	4.43	0.80	4.02	0.76	-	-	-	-	4.12	0.90
	ChatGLM-6B	2.88	0.37	2.92	0.61	3.10	0.60	-	-	-	-	2.97	0.53
	ERNIE Bot	3.48	0.50	3.40	0.82	3.63	0.84	-	-	-	-	3.51	0.72
	SparkDesk	3.05	0.62	3.08	0.53	3.22	0.61	-	-	-	-	3.12	0.58
Score	human experts	4.89	0.18	4.59	0.60	4.61	0.37	4.83	0.30	4.78	0.33	4.74	0.36
	ChatGPT	3.75	0.55	3.39	0.48	4.39	0.35	4.52	0.39	4.29	0.57	4.07	0.47
	GPT-4	4.10	0.78	4.26	0.48	4.46	0.37	4.68	0.35	4.33	0.62	4.37	0.52
	ChatGLM-6B	4.03	0.32	3.44	0.41	4.20	0.33	4.54	0.46	4.02	0.56	4.05	0.42
	ERNIE Bot	4.32	0.28	3.86	0.73	4.46	0.36	4.65	0.43	4.06	1.20	4.27	0.60
	SparkDesk	4.09	0.62	3.56	0.42	4.32	0.31	4.74	0.39	4.39	0.48	4.22	0.45

TABLE III. HUMAN SCORING RESULTS OF FORMAL STYLE GENERATED TEXTS IN LLMS AND GRAND AVERAGE

Indicator	Text Source	Official Document		Political Comment		Academic Literature		News		AVG		Grand AVG	
		AVG	SD	AVG	SD	AVG	SD	AVG	SD	AVG	SD	AVG	SD
Appropriateness	human experts	4.80	0.48	4.83	0.37	4.70	0.46	4.70	0.46	4.76	0.44	4.82	0.40
	ChatGPT	4.57	0.74	4.08	0.80	4.47	0.65	4.45	0.56	4.39	0.69	4.24	0.71
	GPT-4	4.83	0.37	4.33	0.77	4.82	0.43	4.45	0.74	4.61	0.58	4.56	0.64
	ChatGLM-6B	4.45	0.72	4.22	0.69	4.80	0.44	4.45	0.87	4.48	0.68	4.17	0.71
	ERNIE Bot	3.98	1.18	1.42	2.04	4.85	0.51	4.70	0.49	3.74	1.05	3.99	0.92
	SparkDesk	4.80	0.44	2.85	1.74	4.60	0.74	4.37	1.14	4.15	1.01	4.16	0.86
Normativeness	human experts	4.73	0.51	4.80	0.48	4.60	0.61	4.50	0.67	4.66	0.57	4.70	0.52
	ChatGPT	4.03	0.86	4.35	0.68	4.57	0.62	4.15	0.75	4.28	0.73	4.30	0.71
	GPT-4	4.27	0.68	3.93	0.85	4.55	0.69	3.75	0.85	4.13	0.77	4.23	0.75
	ChatGLM-6B	4.23	0.74	4.65	0.57	4.67	0.65	4.28	0.90	4.46	0.71	4.61	0.60
	ERNIE Bot	4.60	0.76	1.58	2.25	4.75	0.47	4.65	0.48	3.90	0.99	4.38	0.79
	SparkDesk	4.70	0.49	3.47	2.08	4.85	0.57	4.48	1.12	4.38	1.07	4.61	0.72
Score	human experts	4.77	0.40	4.82	0.33	4.65	0.49	4.60	0.47	4.71	0.42	4.73	0.39
	ChatGPT	4.30	0.63	4.22	0.57	4.52	0.47	4.30	0.51	4.33	0.55	4.19	0.50
	GPT-4	4.55	0.43	4.13	0.58	4.68	0.41	4.10	0.65	4.37	0.52	4.37	0.52
	ChatGLM-6B	4.34	0.55	4.43	0.47	4.73	0.40	4.37	0.79	4.47	0.55	4.23	0.48
	ERNIE Bot	4.29	0.83	1.50	2.13	4.80	0.37	4.68	0.37	3.82	0.93	4.07	0.74
	SparkDesk	4.75	0.34	3.16	1.87	4.73	0.50	4.43	1.07	4.27	0.95	4.24	0.67

he level of "relatively good", but the scores of literary grace of LLMs except GPT-4 are generally low, with written evaluations such as "bland description", "lack of emotion", "empty content" and so on. Compared with human experts' emotion, thoughts, expressiveness and creativity, LLMs still have a large gap.

D. Stylistic Transformation

There are a large number of sentences in SCCS, from which we screen 463 sets of informal-formal parallel sentences of high quality that fit the context of modern Chinese as a test set, with an average sentence length (in characters) of 12.57. The prompt is as follows:

请将下列文本以更加书面/口语化的方式进行表达。

文本：{text}

(Please present the following text in a more written/spoken way.

Text: {text})

This task adopts the method of human evaluation to appraise the transformation competence of LLMs, using formality/informality, normativeness and consistency as indicators to measure the quality of the transformation results, and the final score is the average value of each indicator. A total of 6 annotators are recruited for this task, all of whom are postgraduates majoring in linguistics. In addition, we add 116 sets of high-quality human transformation results as human experts indicator into LLMs' transformation results at a ratio of 4:1 without informing the annotators. Among them, the human transformation results of informal-formal are taken from the parallel sentences in SCCS, and the formal-informal direction is also manually transformed by two linguistics postgraduates. The formal-informal stylistic transformation specification is basically the same as the construction specification of the SCCS, which is required to serve the modern Chinese style, and the normativeness and semantic consistency of the transformation results are also demanded. The annotating specifications are as follows:

a) Indicators and Scores: The annotators score formality/informality, normativeness and consistency on a scale of 1-5 respectively, and provide written evaluations. Among them, formality/informality focuses on the degree of stylistic change of the transformation result compared with the original text. Normativeness as above. Consistency focuses on the consistency and completeness of semantics before and after transformation.

b) Invalid outputs: If LLM does not generate the transformation result, the original text is not transformed, or the transformation result is English, all indicators are assigned 0 points.

c) Special cases: If the generated result is classical Chinese, the formality is uniformly judged to be 1 point, and other indicators are normally scored. If other redundant information appears in the text, it will be ignored when scoring and only the transformation result will be scored. If multiple transformation results are generated, the final score takes the highest value of multiple transformation results.

d) The misconceptions to be avoided and the requirements of other evaluation items are the same as above.

Finally, the average value and standard deviation of the scoring results of each indicator are utilized to evaluate the stylistic transformation competence of LLMs, and the data are shown in Table IV.

In the inter-transformation of style, GPT-4 shows the most excellent and stable performance, followed by ChatGPT and ERNIE Bot. ChatGLM-6B also obtains better scores, whereas SparkDesk exhibits remarkably poor transformation competence, with numerous instances of untransformed original texts. Overall, except for SparkDesk, each model achieves a two-way balance in stylistic transformation competence, albeit with a noticeable disparity compared to human experts.

Specific to the three indicators, it is found that several LLMs have outperformed human experts in terms of formality scores, highlighting the outstanding ability of LLMs in formal style once again. Of course, since the construction of SCCS is based on *Tongyici Cilin*, the stylistic changes of some parallel sentences mainly focus on lexical stylistic transformation, which may affect the scoring of the formality of human experts to some extent. However, the informality of LLMs, especially the consistency scores, still have a considerable distance from human experts. Similarly, the predominantly written training corpus of LLMs limits their informality, and LLMs suffer from the common problem of hallucination, which is more clearly exposed during the writtenization process. Summarizing the written evaluations, it is found that illogicality, inappropriate metaphors, and lack of fluency in the generated texts continue to be a problem.

TABLE IV. HUMAN SCORING RESULTS OF LLMS STYLISTIC TRANSFORMATION

Indicator	Source	Informal-Formal		Formal-Informal		Grand AVG	
		AVG	SD	AVG	SD	AVG	SD
Formality/Informality	human experts	4.72	0.55	4.81	0.41	4.77	0.48
	ChatGPT	4.75	0.67	4.55	0.75	4.65	0.71
	GPT-4	4.81	0.61	4.65	0.64	4.73	0.62
	ChatGLM-6B	4.32	1.23	4.31	0.87	4.31	1.05
	ERNIE Bot	4.74	0.72	4.54	0.72	4.64	0.72
	SparkDesk	2.03	1.95	0.87	1.57	1.45	1.76
Normativeness	human experts	4.97	0.16	4.99	0.09	4.98	0.13
	ChatGPT	4.83	0.44	4.86	0.49	4.85	0.46
	GPT-4	4.85	0.41	4.93	0.31	4.89	0.36
	ChatGLM-6B	4.58	1.14	4.73	0.73	4.66	0.94
	ERNIE Bot	4.83	0.68	4.94	0.34	4.88	0.51
	SparkDesk	2.67	2.48	1.23	2.15	1.95	2.32
Consistency	human experts	4.85	0.45	4.88	0.35	4.86	0.40
	ChatGPT	4.25	0.98	4.32	0.98	4.29	0.98
	GPT-4	4.37	0.84	4.46	0.78	4.42	0.81
	ChatGLM-6B	3.51	1.47	3.75	1.25	3.63	1.36
	ERNIE Bot	4.18	1.08	4.38	0.85	4.28	0.96
	SparkDesk	2.63	2.45	1.20	2.10	1.91	2.28
Score	human experts	4.85	0.26	4.89	0.20	4.87	0.23
	ChatGPT	4.61	0.43	4.58	0.52	4.60	0.48
	GPT-4	4.68	0.38	4.68	0.37	4.68	0.38
	ChatGLM-6B	4.13	1.08	4.26	0.67	4.20	0.87
	ERNIE Bot	4.58	0.66	4.62	0.43	4.60	0.55
	SparkDesk	2.45	2.28	1.10	1.93	1.77	2.10

E. Analysis of Results

Based on the above statistical results, the scores of human experts and LLMs are calculated and ranked using the F1 average of the stylistic classification results, and the average score of the human scoring results of stylistic generation and stylistic transformation. The comprehensive score is the total score of the three types of tasks (with a score range of [0,11]). The results are shown in Table V.

GPT-4 ranks first in all three types of tasks, with a total score of 9.907. ChatGPT and ERNIE Bot follow closely, and their stylistic competence is relatively close. Both ChatGLM-6B and SparkDesk have obvious shortcomings, the former has weak performance in classification task, while the latter lacks stylistic classification, especially stylistic transformation competence. This also verifies that larger scale models usually perform better on generative tas-

TABLE V. RANKING AND SCORE OF LLMS' CHINESE STYLISTIC COMPETENCE

Ranking	Stylistic Classification	Stylistic Generation	Stylistic Transformation	Comprehensive Performance
0	human experts（1）	human experts（4.727）	human experts（4.871）	human experts（10.598）
1	GPT-4（0.860）	GPT-4（4.367）	GPT-4（4.680）	GPT-4（9.907）
2	ChatGPT（0.798）	SparkDesk（4.240）	ERNIE Bot（4.602）	ChatGPT（9.579）
3	ERNIE Bot（0.690）	ChatGLM-6B（4.234）	ChatGPT（4.595）	ERNIE Bot（9.360）
4	SparkDesk（0.300）	ChatGPT（4.186）	ChatGLM-6B（4.198）	ChatGLM-6B（8.708）
5	ChatGLM-6B（0.276）	ERNIE Bot（4.068）	SparkDesk（1.772）	SparkDesk（6.312）

ks. Of course, different product positioning and development rules of each model (such as the degree of control over sensitive data) will also have a certain impact on its actual generation results.

As mentioned above, the performance of LLMs in some formal styles has become comparable to human experts, which brings more or less impact to numerous industries, accompanied by many risks and fallacies. However, the impact of LLMs must be approached squarely, and should also be scrutinized through an instrumental lens in order to promote the development of technology itself, with a view to obtaining dividends while mitigating the risks of the practice [27]. From the perspective of language resources, on the one hand, language resources are the basis for the development of LLM. More diverse language expression scenarios should be included to achieve a balance between depth and breadth, and provide sufficient high-quality stylistic data resources for LLM to meet complex practical application needs. On the other hand, LLM is an important language resource, which can be regarded as a huge stylistic resource corpus to serve the language life of the country, industry, individual and other levels, including text writing, text polishing, text simplification, accompanying chatting, assisting language teaching, lexicography and so on.

V. CONCLUSION

In this study, we devise and practice a set of performance evaluation schemes for Chinese stylistic competence of LLM, and investigate the Chinese stylistic competence of ChatGPT, etc. in a more comprehensive way from the three tasks of stylistic classification, stylistic generation, and stylistic transformation, finding that LLMs have their own advantages and limitations in different tasks and styles. For LLM, it should be viewed from an instrumental perspective, expanding its stylistic competence with rich and diversified stylistic data resources and technological advances, and at the same time reasonably utilizing and giving full play to its stylistic resources attribute, to make it better serve the language life [27]. In addition, although the present study focuses on evaluating LLMs, it is also an important attempt to evaluate human stylistic competence, which carries reference value for language competence enhancement in the fields of Mandarin education and international Chinese language teaching.

REFERENCES

[1] Jianfang Li, "An Analysis of the Acquisition Process of Style", Contemporary Rhetoric, No.1, pp. 14-15, 1998.

[2] Chunxiang Wu, "Ruminations on the experimental methodology of stylistic competence", Contemporary Rhetoric, No.4, pp. 33-42, 2014.

[3] Yun Zhou, Yongqin Zhang, Jing Zhang, "The Stylistic Competence of CFL Learners", Journal of Yunan Normal University(Teaching and Research on Chinese As A Foreign Language Edition), vol 8, No.1, pp. 49-54, 2010.

[4] Minglong Hu, "Analysis of stylistic errors in the writing of Chinese as a second language learners", Shaanxi Normal University, MA thesis, 2013.

[5] Lin Sheng, "Error Analysis of Stylistic Features in Second Language Acquisition: A Study Based on the Chinese Interlanguage Corpus", Journal of ZheJiang Normal University(Social Sciences), vol 37, No.6, pp. 66-70, 2012.

[6] Yanyan Cheng, "Research of Foreign Students' Colloquial Tendency in Chinese Writing", Jilin University, MA thesis, 2017.

[7] Lei Zhang, "A Study on the Chinese Literary Competence of Advanced Foreign Students", Jinan University, MA thesis, 2018.

[8] Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A. et al., "Benchmarking generalization via in-context instructions on 1,600+ languagetasks," arXiv preprint arXiv:2204.07705, 2022.

[9] Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y. et al., "Agieval: A human-centric benchmark for evaluating foundation models,"arXiv preprint arXiv:2304.06364, 2023.

[10] Zhang, X., Li, C., Zong, Y., Ying, Z., He, L., and Qiu, X, "Evaluating the Performance of Large Language Models on GAOKAO Benchmark,"arXiv preprint arXiv:2305.12474, 2023.

[11] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y. et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," arXiv preprintarXiv:2306.05685, 2023.

[12] Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A., "Orca: Progressive Learning from Complex Explanation Traces of GPT-4," arXivpreprint arXiv:2306.02707, 2023.

[13] Wang, Y., Yu, Z., Zeng, Z., Yang, L., Wang, C., Chen, H. et al., "PandaLM:An Automatic Evaluation Benchmark for LLM Instruction TuningOptimization," arXiv preprint arXiv:2306.05087, 2023.

[14] Amin, Mostafa M., Erik Cambria, and Björn W. Schuller, "Will Affective Computing Emerge From Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT," IEEE Intelligent System, 2023, pp.15-23.

[15] W. Jiao, W. Wang, J. Huang, X. Wang, and Z. Tu, "Is ChatGPT A Good Translator? A Preliminary Study," arXiv, Jan. 31, 2023. doi: 10.48550/arXiv.2301.08745.

[16] Huapin Zhang, Linhan Li, Chunjin Li, "ChatGPT Performance Evaluation on Chinese Language and Risk Measures", Data Analysis and Knowledge Discovery, vol 7, No.3, pp. 16-25, 2023.

[17] R. A. Khan, M. Jawaid, A. R. Khan, and M. Sajjad, "ChatGPT - Reshaping medical education and clinical management," Pakistan Journal of Medical Sciences, vol. 39, no. 2, Feb. 2023.

[18] A. Rao, J. Kim, M. Kamineni, M. Pang, W. Lie, and M. D. Succi, "Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making," medRxiv, Feb. 07, 2023.

[19] S. Shahriar, J. Ramesh, M. Towheed, T. Ameen, A. Sagahyroon, and A. R. Al-Ali, "Narrative Integrated Career Exploration Platform," Frontiers in Education, vol. 7, 2022.

[20] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F. et al., "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education," EdArXiv, Jan. 29, 2023.

[21] Shengli Feng, “On mechanisms of Register System and its grammatical property”, Studies of the Chinese Language, No.5, pp. 400-412+479, 2010.

[22] Qinqing Tai and Gaoqi Rao. “A study on the measurement and classification of Chinese stylistic features,” in Proceedings of the 20th Chinese National Conference on Computational Linguistics, 2021, pp. 398-412.

[23] Guojing Huang, Liwei Zhou, Gaoqi Rao, and Jiaojiao Zang. “Construction of Chinese register classification resources based on Tongyici Cilin”, in Proceedings of the 21st Chinese National Conference on Computational Linguistics, 2022, pp. 431-443.

[24] Shengli Feng, “An Introduction to Chinese Register Grammar”, Beijing:Beijing Language and Culture University Press, 2018.

[25] Jianming Lu, “As a second language teaching, Chinese teaching must pay attention to written language teaching”, Research on Chinese as a Second Language, No.0, pp. 109-113, 2007.

[26] Yi Li, Shike Wang, Dong Yu, Pengyuan Liu, “On the Construction of Linguistic Feature System and Automatic Evaluation for Literary Grace of Chinese Texts”, Applied Linguistics, No.1, pp. 130-144, 2023.

[27] Gaoqi Rao, Xingyu Hu, Zilin Yi, “Governance of Large Language Models from the Perspective of Language Resources”, Chinese Journal of Language Policy and Planning, vol 8, No.4, pp. 19-29, 2023.