Holistic Evaluation for LLM's Capability in Human-level Writing using Tree of Writing

Anonymous ACL submission

Abstract

Evaluating the writing capabilities of large language models (LLMs) remains a significant challenge due to the multidimensional nature of writing skills and the limitations of existing metrics. LLM's performance in thousand-words level and open-ended writing is inadequately assessed by traditional reference-based metrics or modern LLM-as-a-judge methods. We propose Tree-of-Writing (**ToW**), aiming to solve the implicit inconsistency often found when LLM-as-a-judge aggregates all sub-features in text evaluation. **ToW** incorporates a tree-structured workflow by explicitly modeling the aggregation weights of sub-features. We also present HOWTOBENCH, a large-scale Chinese writing benchmark encompassing $\mathbf{12}$ genres and $\mathbf{1302}$ instructions across three task categories: contextual **completion**, outline-**guided** writing, and **open**-ended generation. **ToW** successfully mitigates the biases, achieving a $\mathbf{0.93}$ Pearson correlation with human judgments. Furthermore, we detect that both overlap-based text generation metrics and popular LLM-as-a-judge practices are vulnerable to textual disturbances, while **ToW** are robust to them. We also uncover a negative correlation between input length and content related scores in **Guide** task, showcasing that LLM writings cannot be simply improved by input-side information piling.

1 Introduction

The advances of large language models (LLMs) (Ouyang et al., 2022; Rafailov et al., 2024) have revolutionized the field of natural language processing, enabling breakthroughs in tasks like text summarization (Basyal and Sanghvi, 2023), machine translation (Zhu et al., 2024), conversational agents (OpenAI, 2022; Team-GLM, 2024; Gemini-Team, 2024a), and creative writing (Mostafazadeh et al., 2016; Fan et al., 2018). Despite their promising performance, auto-evaluating LLM-generated text remains a critical challenge particularly in complex, open-ended writing scenarios (Köksal et al., 2024; Yang et al., 2024; Khatun and Brown, 2024).

The ability to generate nuanced and contextually appropriate writing depends heavily on handling implicit requirements, a challenge faced by both humans and LLMs. Existing evaluation methods for LLMs' writing skills predominantly focus on explicit instruction fulfillment (Liu et al., 2024; Kim et al., 2024b; Zhu et al., 2023; Wu et al., 2025), i.e., whether the content meets the requirements; but this narrow focus, akin to a *"mimicking game"*, overlooks LLMs' ability to craft complex, nuanced texts like fictional narratives or persuasive speeches where the intents behind the requirements are much more implicit but directly drive the requirement.

Current approaches (Kim et al., 2024a; Zhu et al., 2023; Wu et al., 2025) often rely on descriptions of evaluation criteria as instructions to the LLM-evaluator, requiring LLMs to provide sub-scores (e.g., fluency, consistency, instruction-following) leading to a final assessment. However, simply averaging the sub-scores is not necessarily an accurate reflection of overall quality, and LLM auto-planned negotiations between rubrics (Wu et al., 2025) result in inconsistent and opaque assessment in multiple runs and queries. This misalignment with evaluation guidelines, which we quote as '*Negotiation Inconsistency*', results in unreliable and opaque assessments, undermining the credibility of LLM-as-a-judge in such tasks.

To address the challenge of *Negotiation Inconsistency* in writing assessment, we propose the **Tree-of-Writing** (**ToW**) framework, which simulates the human decision-making process. **ToW** operates on a well-structured tree, which treats key evaluation aspects—such as language, logic, and plot—as leaf nodes. For each writing instruction, an LLM-negotiator designs the aggregation plan based on genre, task type and other requirements.

Through a depth-first traversal on the plan, corresponding sub-score expert agents are activated to score each aspect. **ToW** achieves a transparent and reproducible assessment for nuanced writings.

Distinct from existing benchmarks (Liu et al., 2024; Zhu et al., 2023; Kim et al., 2024b; Wu et al., 2025) which all treat writing as a *"mimicking game"*, we propose HOWTOBENCH, a large-scale benchmark designed to evaluate LLMs' writing abilities through three carefully designed task formats—**Completion, Guide, Open**—reflecting varying levels of provided context. HOWTOBENCH spans **12** genres with **1302** writing instructions, covering both creative and functional tasks. The dataset is curated from expert-written sources, highlighting the goal to emulate human-professional writing. The final pass-rate for dataset quality check by human experts is 96.85%.

To validate the effectiveness of **ToW**, we conducted large-scale evaluations on the writings generated by 10 flagship LLMs, including Gemini-2.0-flash (Gemini-Team, 2024b), GPT-4o-1120/o3-mini (OpenAI, 2024), Claude-3.5-Sonnet (Claude-Team, 2024b) and DeepSeek-R1/V3 (DeepSeek-AI, 2025, 2024). Our framework demonstrates strong alignment with human preferences, achieving a Pearson correlation up to **0.93** when comparing system rankings with human-annotated rankings for all LLMs-generated writings.

Through our evaluation, we found that some LLMs such as GPT-series demonstrate strong performance in a rich-context setting (**Completion**) but dropped drastically when the input information is limited . In our analysis to all generated writings, we concluded that though positive correlation is found between input and output length, it is significant that longer inputs and outputs are related with lower overall assessment, indicating the challenges of the tasks other than length trick. Furthermore, most metrics, including the LLM-as-a-judge practices, are vulnerable to contextual fallacies, such as repetition, in proper styles.

To the best of our knowledge, we are the first to explore the assessment of LLMs' capabilities in human-level writing with elaborately designed instructions beyond the instruction-following view.

2 Related Work

2.1 Benchmarking LLM Writing

Prior research on LLM writing evaluation has predominantly focused on creative story generation, emphasizing fluency and coherence through datasets like RocStories (Mostafazadeh et al., 2016) and metrics like OpenMEVA (Guan et al., 2021). While these works highlight narrative quality, their scope is limited to predefined genres (e.g., fiction) and narrow evaluation dimensions. Recent benchmarks for general text generation assess instruction-following (Zheng et al., 2023; Liu et al., 2024), lexical and coherence (Zhang et al., 2024a,b) or domain expertise (Liang et al., 2023), yet they inadequately address the open-ended nature of writing tasks. Very recent work (Wu et al., 2025) seek an instruction-following way for writing evaluation. For instance, reference-based metrics (Deutsch et al., 2022) prioritize structural conformity over creative divergence, while existing LLM-as-a-judge methods struggle with genre-specific stylistic nuances.

HOWTOBENCH advances this line of research by (1) expanding evaluation to **12** diverse genres beyond fiction, (2) disentangling format, content, and subjective impression during evaluation, and (3) explicitly addressing the tension between instruction adherence and creative openness through three task categories.

2.2 LLM-based Evaluation

Recent advances in LLM-based evaluation leverage proprietary models for automated scoring through prompt engineering (Zheng et al., 2023; Liu et al., 2023) or supervised training on human annotations (Wang et al., 2024b; Ke et al., 2024). These methods surpass traditional metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) in efficiency and human correlation, particularly for constrained tasks (e.g., summarization). However, their reliability diminishes in open-ended writing evaluation: verbosity bias (Zheng et al., 2023), positional bias (Wang et al., 2024a), and rubric dependency (Ke et al., 2024; Kim et al., 2024a) limit generalizability across genres. On the other hands, attempts (Wu et al., 2025) where LLM automatically plans the evaluation criteria and rubrics emerged, yet its robustness are ignored.

Comparison of our work to previous works are listed in Table 1.

3 Evaluation Methodology

3.1 Tree-of-Writing Mechanism

We introduce Tree-of-Writing (**ToW**), aiming to solve the hierarchal judgment nature of writing.

	Size	#Tasks	Lang	Ref Source	Domain	Open	IF	Dims	Metric
Fan et al. (2018)	10k+	1	EN	Reddit writing prompts	100-200 Story	Yes	No	No	BLEU, ROUGE
Kim et al. (2024b)	770	9	EN	LLM generated mainly	IF & Reasoning & Safety	No	Yes	Rubric	LLM metric
Guan et al. (2022)	729	4	CN	Online stories	Story	Yes	No	No	BLEU, DIST
Zheng et al. (2023)	10	1	EN	Self Constructed	Functional Writing	No	Yes	General	LLM-as-a-judge
Liu et al. (2024)	75	4	CN	LLM generated	Text Generation	No	Yes	Rubric	LLM-as-a-judge
Wu et al. (2025)	1239	6	CN/EN	LLM generated with human refine	IF-style writing	No	Yes	Auto-Plan	LLM-as-a-judge
HOWTOBENCH (Ours)	1302	3x12	CN	Professional	Creative & Functional Writing	Yes	Yes	Text Features	**ToW**

Table 1: Differences between our works from previous advances in natural language generation and instruction following fields. **Lang** shorts for language. **Ref** shorts for reference. **EN** shorts for English and **CN** shorts for Chinese. **IF** shorts for instruction following.

We refer to Figure 1 for better illustration, where a poetry is being judged with tree-of-writing. Let R be the root of the tree of the poetry. For each genre, three children are derived from the root: V_C the content child, V_F the format child, and V_I the impression child. Each child are linked to the root with weighted edge E_C, E_F, E_I. For V_C and V_F, there are more atomic leaf children L_i, which are also weighted connected to their parent with $E_{V_{\text{Parent}(L_i)}L_i}$. V_I do not have children and therefore is a leaf child of the tree. $\text{Parent}(\cdot)$ refers to the parent function which returns the parent of the variable node.

The scoring of the poet is calculated with a DFS of the tree:

$$\text{Score}(V_C) = \sum_{L_i \in \text{Child}(V_C)} w_{E_{V_C L_i}} \text{Score}(L_i)$$

$$\text{Score}(V_F) = \sum_{L_i \in \text{Child}(V_F)} w_{E_{V_F L_i}} \text{Score}(L_i)$$

$$\text{Score}(R) = \sum_{j \in \{C,F,I\}} w_{E_j} \text{Score}(V_j)$$

$\text{Child}(\cdot)$ refers the children function which returns the childrens of the variable node.

3.2 Scoring Function

There is a key issue in the implementation of the above $\text{Score}(\cdot)$ function: for different types of nodes, we have used different methods to implement the $\text{Score}(\cdot)$ function.

For the **Content** nodes V_C, each leaf node corresponds to a specific trait. We implemented them using a combination of rubric with reference approach. Formally speaking, an LLM is utilized to assign a score between 1 and 10 to the current leaf node. The corresponding descriptions are provided in Table 7.

For the **Format** nodes, we adopt a hybrid approach combining rule-based and LLM-based methods. The scoring function follows a 0/5/10 step function. For the *Plots & Structure* and *Paragraphing* nodes, an LLM-based judge evaluates whether the content's structure and level of detail are appropriate. For the *Formatting* leaf nodes, a regex-based approach is detect whether the titling are appropriate or violating the hierarchical relations. The detailed rules are outlined in Table 8. Detail implementation of Regex is attached in Appendix M.

3.3 Edge Weighting

We adopt a method that assigns weights for leaf nodes and performs a weighted average for intermediate nodes based on the number of leaf nodes.

For leaf nodes, we use an *explicit* edge weighting approach. First, an LLM edge-weight donar determines the edge weights based on instruction $\mathcal{I}$, ensuring all weights are between -1 and 1 and sum up to 1:

$$(w_{E_{V_C L_1}}, \cdots, w_{E_{V_C L_n}})^i = \text{J}_W(\mathcal{I}^i)$$

$$\sum_{k=1}^{n} w_{E_{V_C L_k}} = 1, w_{E_{V_C L_k}} \in (-1, 1)$$

Once the leaf node scores are determined, we aggregate them using these weights. It avoids inconsistencies in the *implicit* aggregation strategy employed by the LLM, such as randomly choosing between averaging or favoring certain dimensions within the same instruction. Moreover, this approach enhances the interpretability of the evaluation results, facilitating further analysis. The implementation prompts of this part is attached in Appendix L.1.

For the aggregation of $\text{Score}(V_C)$, $\text{Score}(V_F)$ and $\text{Score}(V_I)$, we use an averaging method based on the number of leaf nodes. This allows task like completion, which may lack a format dimension, to be integrated into a consistent evaluation framework. It also offers advantages when extending tasks types.

4 HOWTOBENCH

To holistically evaluate the capabilities of LLMs in generating human-level writings, we devel-

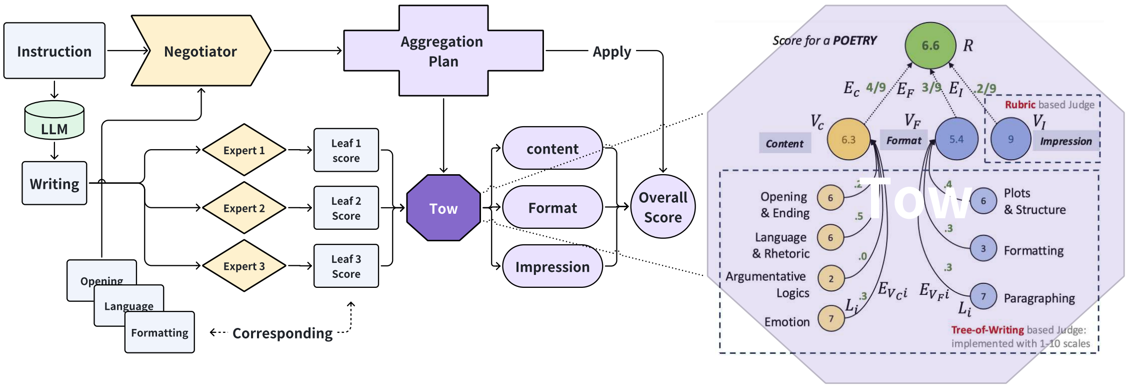

Figure 1: Overview of the evaluation framework incorporating the **ToW**.

oped HOWTOBENCH, which is designed across a range of writing genres, in form of 3 single-round writing mode (**Completion, Guide, Open**) tasks. HOWTOBENCH is characterized for its high-quality expert written reference without contamination from AI-generated content.

4.1 Task Definition

LLM-based writing tasks are formalized within an input-output framework.

Writing instruction $\mathcal{I}$: lists the requirements for the writing task. It also includes a one-sentence summary of desired writing.

Grounding information $\mathcal{G}$: encompassing supplementary details such as formatting requirements, narrative or plot constraints, stylistic directives, or no additional context.

Human reference $\mathcal{R}$: a curated and de-noised high-quality reference to the task. It plays an important role in the evaluation.

Based on these inputs, the LLM generates an output writing:

$$\mathcal{W} = \text{LLM}(\mathcal{I}, \mathcal{G})$$

which is for reconstructing the human-level quality from the summary in the instruction, and the generation are evaluated in the content and format.

4.2 Data Source: Crawling

We crawled a large set of high-quality, publicly licensed human-written texts from specialized literary and writing guide websites: **CN Writer**, **PW4ES**, **SeptES**, **ZJPub**, **Officials**. They are introduced in detail in **Appendix** A.1. These texts address the aforementioned issues and are all written by human writers or experts.

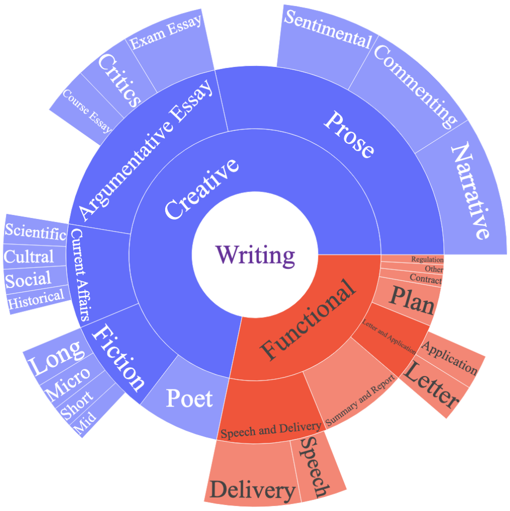

Figure 2: Hierarchal taxonomy of HOWTOBENCH showing the major categories.

4.3 Reference: Categorizing and Filtering

We use a category classifier to classify the above mentioned crawled text T into each writing genres $c = \text{Cls}(T)$. Specifically, we implement it with a prompted LLM. We prompt a GPT-4o-1120 with the following prompts in **Appendix** D. We employ three human experts at LLM writing[1] to manually check the GPT-classified tags. For all 1302 prompts, GPT-4o reached 98.6% accuracy. The three human experts have manually revised those wrongly classified.

The concerned writing genres are: fiction, poet, prose, essay, argumentatives, reports, summaries, letters, speeches, deliveries, plans, contracts, officials. Further introductions for each genres are listed in **Appendix** A.2.

Furthermore, to ensure bench data quality, we

[1] Master degree in humanities, journalism, finance respectively with two working experience in LLM industry.

harness another LLM as a filter to eliminate the low quality texts from the crawled data. Specifically, we get a overall quality score from 1 to 5 $s = \text{Filter}(T)$ from the LLM. Higher score means better quality. We implement it with Claude-3-5-sonnet-20241022. We prompt the Claude with 13 genre-specific rubrics. We attach the prompt for fiction in **Appendix** E as an example. We show the score distribution for the mentioned websites in Section 4.2 in Table 11. Most of the texts got scores in 3 to 5. We set the threshold score at 4 and discard all examples scored less than threshold.

4.4 Task Design: Progressive Difficulty Levels

To evaluate the key writing capabilities of LLMs we design completion, guided writing, and open writing with progressive difficulty. As the constraints and prompts for writing gradually decrease from Level I to Level III, it becomes increasingly challenging to plan and expand the writing with less input information. We provide examples for each task in Appendix N.

Level I: Completion: This task evaluates the LLM's ability to complete the context of unfinished text. In this task, key portions of a text are omitted. The instruction $\mathcal{I}$ requests the LLM to complete the omitted texts in the grounding information $\mathcal{G}$.

Level II: Guided Writing: This task measures the LLM's ability to expand and generate text based on an outline. The instruction $\mathcal{I}$ directs the LLM to adhere to the grounding information $\mathcal{G}$, which contains a predetermined theme and genre.

Level III: Open Writing: This task evaluates the LLM's ability to freely elaborate on a given topic. The instruction $\mathcal{I}$ only requires the genre and discloses the topic, plot or argument within one sentence. There are no given grounding information, i.e. $\mathcal{G} = \varnothing$.

4.5 Instruction: Reverse Construction

We construct the instruction $\mathcal{I}$ and grounding information $\mathcal{G}$ based on the high quality references. We refer to this process as back-construction, for its similarity to back-translation.

Instruction Template for completion

Input: Genre
Please fill in the blanks in the following {genre}, marked with [fill in the blank] signs. You should comprehensively consider the context and ensure the completion quality.

	Comp	Guide	Open	Total
#Creative	379	277	282	938
Instr Len	44.02	88.82	89.29	70.86
Info Len	2016.02	318.48	-	1299.22
Ref Len	431.37	1607.52	1726.05	1167.93
#Functional	-	179	185	364
Instr Len	-	85.68	91.83	88.80
Info Len	-	467.39	-	467.39
Ref Len f	-	1335.23	1373.91	1354.89

Table 2: Statistics of HowToBench.

For **Completion**, we enrolled human annotators manually remove portions of text from human-written content with paragraphs as the smallest unit of granularity. The maximum removal of paragraphs is limited to 10. We set the the incomplete human writing as $\mathcal{G}$ and the removal as the reference $\mathcal{R}$. The instruction $\mathcal{I}$ is composed using the template below.

Instruction Template in guided/open writing

Inputs:Genre,Topic,Summary,Word counts
Please write a {genre} about {Topic}. {summary}. You should write in approximately {word counts}.

For **Guide** and **Open**, we utilize a LLM as the back constructor. Formally, it goes:

$$(S, T, \mathcal{G}) = \text{BackConstruct}(\mathcal{R})$$

where S and T refer to the summary of the original sentence and the theme consisting of no more than five words, respectively. S and T are filled in the following template to construct $\mathcal{I}$. Besides, the back-constructor is assigned specific traits of a the genre, and it needs to provide descriptions of writing requirements based on these traits, depending on $\mathcal{R}$. All the traits information is then composed in $\mathcal{G}$. We implement the back-constructor with Gemini-2.0-Flash. We also prompt it with one-shot in context example. The prompt for genre fiction is attached to **Appendix** F.

4.6 Quality Assurance

The initial curation for instruction and information are synthetic. We inspect and manually make revisions for all the HowToBench data. Specifically, we again enroll the three experts described in Section 4.3 to determine the quality of the pairs.

For each single pair, $\mathcal{I}$ and $\mathcal{G}$ is firstly inspected in clarify, relatedness to human reference writing, natural expression. They are instructed to the guideline in Appendix G to revise if the case is not qualified. Further, we enhanced the quality for R in a pairwise picking. For all instructions, we made inference on GPT-4o-1120, GLM-4-plus, Gemini-2.0-Flash. Then we arrange them in $\{\mathcal{I}, (\mathcal{G}), \mathcal{R}, \mathcal{W}_{\text{GPT}}, \mathcal{W}_{\text{GLM}}, \mathcal{W}_{\text{Gemini}}\}$ pairs. Human experts then selected the best out of four writings according to the guideline in **Appendix** G. Each pair is judged by two random experts and reached 96.7% agreement rate. Those not agreed by two experts were then determined by the third expert based on previous judgments. $137(10.5\%)$ out of 1302 original human writings were not picked as the best out of the four writings. We substitute them with the one that experts picked.

Then for $\mathcal{I}, \mathcal{G}, \mathcal{R}$ together, personal information, unsafe contents and noises such as advertisement are either removed or revised to desensitized form. During this process, human annotators are assisted with a Detector implemented with Deepseek-R1. The overall un-qualification rate is $41/1302$.

We list the statistic of HOWTOBENCH in Table 2. The length is measured in Chinese characters. In all, the dataset instructions are strict, clear for evaluation, while the reference are high quality to represent the excellence in different writing genres.

5 Experiment

5.1 Baselines and Metrics

We compared two sets of methodologies: auto-metrics including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BLEU-rt (Sellam et al., 2020), LLM-as-a-judge (Zheng et al., 2023) and its derivatives. Among all LLM-as-evaluator derivatives, two typical practices are adopted: **Auto-Planning** and **Elaborated Rubrics**. **Auto-Planning** means the LLM evaluator plans all the subdomains and the aggregation plans together, then produces the scores on its own. **Elaborated Rubrics** means the LLM-evaluator adopts the carefully curated evaluation prompts, which derives directly from human annotation practices. For all genres, their evaluation prompts can be found in Appendix H.

We tested the method on GPT-4o-2024-11-20 (OpenAI, 2024), Gemini-2.0-flash, Deepseek-R1 (DeepSeek-AI, 2025), Deepseek-V3 (DeepSeek-AI, 2024), Doubao-pro-32k (Bytedance-Team, 2024), GLM-4-plus-250111 (GLM-Team, 2024), Claude-3-5-sonnet-20241022 (Anthropic, 2024), Claude-3-haiku-20240307 (Claude-Team, 2024a), Qwen-plus (Qwen-Team, 2025). We conducted experiments on GPT-4o as the base-LLM for all LLM-as-a-judge related methods.

5.2 Meta Evaluation

We release **MetaEditor** as the meta-evaluation dataset for the evaluation design for such writing tasks. **MetaEditor** consists of human ratings on LLM-generated writings in HOWTOBENCH.

We select 221 $(67, 83, 71)$ instructions out of all 1302 prompts for **Completion, Guide, Open**. All genres are randomly and evenly covered. Each instruction is attached with 9 LLM generated writings from Table 4. We hired 36 experts in writing, whose information can be found in Appendix I. We provide writing guideline training for all of the experts and ask them to follow the three annotation guidelines in the Appendix J to score the LLM writings on a scale of 1 to 5. We ensure that the nine writings under the same instruction are graded by the same annotator. For the same LLM writing, we employ two annotators for cross-validation. The overall Inter-Annotator Agreement is 0.71 using Cohen's Kappa and 0.87 using Pearson Correlation, demonstrating a high level of human consistency. We merge the two scores into one by average them, to maintain the diversity of human judgments.[2]

5.3 Results

From the assessment for evaluation methods results in Table 3, we show that **ToW** achieves a 0.93 Pearson and Spearman correlation over all tasks in HOWTOBENCH. Comparing the significance between BLEU and ROUGE-L indicates that the evaluation task, expected evaluation is not based on the recall rate of target, but on precision. This also reflects the design intention to avoid overly depending on specific and explicit requirements. BLEU-rt is a model-based metric showed an random results, indicating that weak base model-based methods are not reliable compared to rule (overlap) based ones. Auto Planning baseline also showed random results in **Completion, Guide**, suggesting its limitation for evaluating tasks with plenty guidance.

In Table 4, we list the performance of LLMs under **ToW**. We categorized the LLMs into three sectors according to their capabilities. It is worth noting that **Completion, Guide, Open** clearly dis-

[2] For all the data to be opensourced, we enroll five experts who have gained the highest agreement with other annotators from the whole annotation process, and instruct them to re-check all the annotations.

	Comp			Guide			Open			ALL		
	ρ	τ	σ	ρ	τ	σ	ρ	τ	σ	ρ	τ	σ
BLEU-1	0.85**	**0.67****	**0.80****	0.65**	0.54**	0.69**	0.70**	0.50**	0.62**	0.75**	0.56**	0.72**
BLEU-rt	0.19	0.06	0.15	-0.25	-0.20	-0.19	-0.45	-0.22	-0.27	-0.19	-0.11	-0.20
ROUGE-L	<u>0.87</u>**	**0.67****	0.75**	0.06	0.14	0.20	0.46	0.22	0.32	0.46	0.06	0.17
Prompt-driven Rubric	**0.88****	<u>0.65</u>**	0.72**	<u>0.79</u>**	<u>0.70</u>**	<u>0.85</u>**	**0.89****	<u>0.72</u>**	<u>0.83</u>**	<u>0.89</u>**	<u>0.61</u>**	<u>0.80</u>**
Auto Planning	0.69**	0.39	0.47	0.46	0.29	0.39	0.71**	0.50*	0.62*	0.79**	0.50*	0.63*
ToW	<u>0.87</u>**	**0.67****	<u>0.78</u>**	**0.85****	**0.76****	**0.89****	**0.89****	**0.78****	**0.88****	**0.93****	**0.83****	**0.93****

Table 3: Assessment for evaluation methods and frameworks. System level Pearson correlation (ρ), Kendall rank correlation τ and Spearman rank correlation σ are calculated. ** marks the p < 0.05 significance and * marks the p < 0.1 significance.

	AVG	DS-R1	o3-mini	4o	CL-35-S	Gemini	DS-V3	DB	GLM	CL-3-H	LM
Completion		6.10	6.16	6.60	5.55	5.43	5.44	5.58	5.19	5.12	4.36
Guide		6.15	5.80	5.61	5.76	5.53	5.52	5.24	5.51	5.08	4.89
Open		6.06	5.69	5.36	5.43	5.33	5.31	5.14	5.28	4.85	4.47
Argumentatvies	5.68	6.24	6.08	6.23	5.73	5.54	5.61	5.74	5.69	5.16	4.77
Comment	5.48	5.95	6.02	5.98	5.54	5.36	5.53	5.30	5.36	5.10	4.65
Poem	5.40	6.00	5.81	6.34	5.41	5.42	5.60	5.15	5.47	4.58	4.20
Prose	5.32	6.25	5.76	5.75	5.49	5.35	5.16	5.06	5.13	4.89	4.33
Fiction	5.07	6.08	5.36	5.37	5.32	5.23	4.82	4.84	4.89	4.53	4.25
Letters	6.02	6.38	6.11	6.13	6.08	6.12	6.18	6.07	6.05	5.47	5.64
Others	5.97	6.33	6.05	6.30	5.91	6.00	6.02	6.42	5.94	5.63	5.12
Speech	5.60	6.01	5.94	5.64	5.80	5.61	5.74	5.66	5.54	5.28	4.83
Report	5.42	5.90	6.00	5.29	5.82	5.26	5.55	5.18	5.11	5.30	4.81
Contract	5.17	5.52	5.80	4.97	5.11	5.08	5.33	5.24	5.18	5.06	4.37
Plan	5.03	5.44	5.75	5.02	4.97	4.94	5.11	5.23	4.83	4.78	4.26
Regulation	4.90	5.31	5.13	4.66	4.91	5.07	4.87	5.07	4.69	4.59	4.72
All		6.10	5.86	5.81	5.58	5.43	5.42	5.34	5.34	5.01	4.59

Table 4: **Bench scores** genre-wisely. For model abbreviations, **DS-R1** refers to Deepseek-R1, **o3-mini** refers to GPT-4-o3-mini-2025-01-31, **4o** refers to GPT-4o-1120, **CL-3.5-S** refers to Claude-3-5-sonnet-20241022, **Gemini** refers to Gemini-2.0-flash, **DS-V3** refers to Deepseek-V3, **GLM** refers to GLM-4-Plus-250111, **DB** refers to Doubao-pro-241225, **CL-3-H** refers to Claude-3-haiku-20240307, **LM-3.3** refers to Llama-3.3-70B-Instruct.

criminate the LLMs that specialize in instruction following while fall behind in open writing, such as GPT-4 series with 7.6%, 18.8% performance drop. Such findings strengthen our claim that the capabilities required for human level writing goes far beyond the instruction following *'mimic game'*.

5.4 Edge Weight Distribution for Content

We analyzed the edge weights assigned by the negotiator to the four leaf nodes under the content node V_C. As illustrated in Figure 3, these weights differ significantly across genres. Interestingly, we observed that the weights for '*logics*' exhibited notable variation within most genres. Additionally, a consistent pattern emerged: the weights for opening-ending remained stable at approximately 10% across all genres. However, across all genres, the edge weights are not evenly distributed among the four leaf nodes. Full plots for all genres can be found in Figure 6 in appendix.

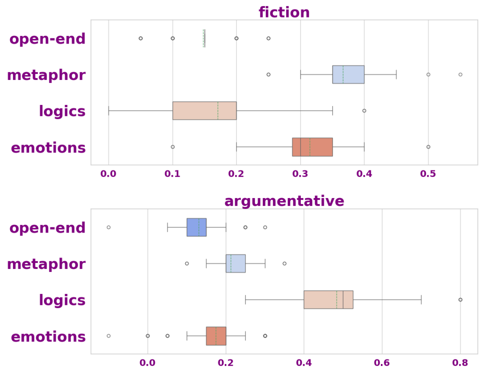

Figure 3: Edge weight distribution on fiction, argumentative. The wider is the box horizontally, the more varied is the corresponding weight within the genre.

6 Analysis

6.1 *Mimic Game*: Longer is NOT Better

We evaluate the impact of input quantity to the LLMs on the writing performance of models. All the writing outputs generated by LLMs are catego-

	Comp	**Guide**	**Open**
Input - Output	0.24**	0.32**	0.25**
Input - Overall	-0.01	-0.44**	-0.15**
Input - Content	-0.01	-0.44**	-0.11**
Input - Format	N/A	-0.16**	0.00*
Output - Overall	0.38**	-0.18**	-0.12**
Output - Content	0.38**	-0.09**	-0.08**
Output - Format	N/A	-0.16**	-0.09**

Table 5: Pearson Correlation between input length, output length and final scores. ** marks the $p < 0.05$ significance and * marks the $p < 0.1$ significance.

	Init.	**Drop**	**Rep**	**To C**	**To L**	**To O**	**To P**
ToW	5.41	-0.36	-0.49	-0.30	-0.31	-0.97	-0.62
Tow-Content	5.82	-0.34	-0.48	-0.17	-0.10	-1.12	-0.36
Tow-Format	5.77	-0.58	-0.81	-0.69	-0.65	-0.74	-1.12
Tow-Impression	6.76	-0.24	-0.30	-0.14	-0.36	-1.52	-0.70
Auto-planning	6.82	-0.06	-0.30	**0.08**	**0.04**	**0.20**	**0.82**
BLEU	24.66	-7.27	**4.23**	**0.97**	**1.21**	-1.56	-8.50
BLEU-rt	37.43	-2.07	-0.35	-2.37	**1.55**	**3.20**	**1.91**

Table 6: Robustness test of frameworks and metrics on common disturbances. **Init.** shorts for initial writing, **Rep** shorts for repetition, **To C/L/O/P** shorts for converting to comment, letter, official, poem. All scores are the results of subtracting the initial score on the left, with a negative sign indicating values lower than the initial score. The **bold red** font indicates the undesired changes.

rized according to **Completion**, **Guide** and **Open**. We conduct correlation analysis and linear regression on the relationships among input length, output length, and final scores, arriving at the results shown in the Figure 5 and Table 5.

There is a significant positive correlation between output length and input length, which is consistent with previous research findings. For the **Guide** and **Open**, we perform linear fitting on the generation results of all models. The slopes are 1.4 and 6.1, respectively, indicating the input tokens conversion ratio to the output.

However, we find that on both **Guide** and **Open** tasks, regardless of **input** or **output**, the final scores exhibit a significant negative correlation with length. This differs from previous understandings where LLM evaluators were thought to favor verbosity. Additionally, we explain that providing more input does not necessarily induce better performance. LLMs are unable to rely on piling up input information to produce high-quality, nuanced writings. This is particularly evident in the Content and Overall scores for Guide tasks, where a correlation of -0.44 was observed.

We leave further discussions to the Appendix, such as different base-LLM evaluators (Appendix B.1), the comparison between reference-based and reference-free LLM judgment (Appendix B.2), between human-originated reference and LLM-originated reference (Appendix B.3).

6.2 *Negotiation Inconsistency* Pro: Robustness

Currently, metrics robustness has aroused community concerns, since *reward hacking* (Skalse et al., 2025) are often encountered in practice. We handle another experiment to validate the **ToW**'s robustness against common text disturbances.

We randomly pick 50 generation samples from LLMs presented in Table 4 (5 for each). We apply the following 3 disturbances to the generated writings following (Guan et al., 2021): (1) **Drop**: randomly drop at most 3 paragraphs or sentences. (2) **Repeat**: repeat at most 3 paragraphs in the original writing at different positions. (3) **Transfer**: convert the writing genre to another different genres. In practice, we pick comment, letter, official, poem as the target genres. We examine **ToW**, auto-planning LLM-evaluator, BLEU, BLEU-rt metrics and show in Table 6.

Through the results, we can find that **ToW** are responding to all the disturbances with score decrement. However, auto-planning, BLEU, BLEU-rt metrics are vulnerable to these interferes, indicating their limitations, which might introduce structures for bypassing designed assessment.

7 Conclusion

This work addresses the issue of "*negotiation inconsistency*" in LLM-as-judges assessment of LLM capability of human-level writing. We find that LLM-as-a-judge is unable to autonomously combine evaluation dimensions through direct prompt engineering or auto-planning. By introducing **ToW**, we enable LLMs to explicitly distinguish the relative importance of sub-dimensions. Overall assessments conducted via tree traversal demonstrates a **0.93** correlation with human judgments on writing tasks. Furthermore, we tackled the bias toward writing assessment present in previous work by designing three task formats—**Completion, Guide**, and **Open**—thus providing a more comprehensive and opaque evaluation of LLMs' human-level writing abilities. Our experiments also explore the relationships between input and output information, and reveal that many existing metrics and practices are unable to detect minor perturbations in simple writing samples.

Limitations

First, although HOWTOBENCH spans 12 genres, its evaluation of writing ability operates at a genre-category level rather than addressing granular subgenres or specialized stylistic variations within each genre. This leaves fine-grained distinctions in domain-specific writing proficiency unexplored.

Second, the evaluation focuses on single-round generation and excludes iterative refinement processes. Methodologies involving self-critique, multi-round human-AI collaboration, or dynamic feedback integration—critical for real-world writing workflows—remain unexplored. This restricts insights into how LLMs adapt to evolving user requirements or contextual adjustments. We leave this scope for future explorations.

Finally, we did not test the scalability of the **ToW** approach, particularly with respect to the correlation between selected dimensions and the feasibility of adding new leaf nodes. Due to the current lack of a comprehensive task framework in the domain of complex text, we adopted a relatively conservative Writing Tree modeling approach.

References

Anthropic. 2024. Claude-3.5-sonnet.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. Preprint, arXiv:2310.10449.

Bytedance-Team. 2024. Doubao-pro-32k blog.

Claude-Team. 2024a. Claude-3.5-haiku blog.

Claude-Team. 2024b. Claude-3.5-sonnet blog.

DeepSeek-AI. 2024. Deepseek-v3 technical report. Preprint, arXiv:2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. Preprint, arXiv:2501.12948.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Gemini-Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Preprint, arXiv:2403.05530.

Gemini-Team. 2024b. Gemini-2.0 blog.

GLM-Team. 2024. Glm-4-plus blog.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2022. LOT: A story-centric benchmark for evaluating Chinese long text understanding and generation. Transactions of the Association for Computational Linguistics, 10:434–451.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6394–6407, Online. Association for Computational Linguistics.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CritiqueLLM: Towards an informative critique generation model for evaluation of large language model generation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13034–13054, Bangkok, Thailand. Association for Computational Linguistics.

Aisha Khatun and Daniel G. Brown. 2024. Assessing language models' worldview for fiction generation. Preprint, arXiv:2408.07904.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In The Twelfth International Conference on Learning Representations.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. Preprint, arXiv:2406.05761.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schuetze. 2024. LongForm: Effective instruction tuning with reverse instructions. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7056–7078, Miami, Florida, USA. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. Transactions on Machine Learning Research. Featured Certification, Expert Certification.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024. AlignBench: Benchmarking Chinese alignment of large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11621–11640, Bangkok, Thailand. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2511–2522, Singapore. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 839–849, San Diego, California. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2024. Gpt-4o blog.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Qwen-Team. 2025. Qwen2.5 technical report. Preprint, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. Preprint, arXiv:2305.18290.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7881–7892, Online. Association for Computational Linguistics.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2025. Defining and characterizing reward hacking. Preprint, arXiv:2209.13085.

Team-GLM. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. Preprint, arXiv:2406.12793.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In The Twelfth International Conference on Learning Representations.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuren Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. Writingbench: A comprehensive benchmark for generative writing. Preprint, arXiv:2503.05244.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. 2024. Seed-story: Multimodal long story generation with large language model. Preprint, arXiv:2407.08683.

Xuanming Zhang, Zixun Chen, and Zhou Yu. 2024a. ProLex: A benchmark for language proficiency-oriented lexical substitution. In Findings of the Association for Computational Linguistics: ACL 2024, pages 8475–8493, Bangkok, Thailand. Association for Computational Linguistics.

Xuanming Zhang, Anthony Diaz, Zixun Chen, Qingyang Wu, Kun Qian, Erik Voss, and Zhou Yu. 2024b. DECOR: Improving coherence in L2 English writing with a novel benchmark for incoherence detection, reasoning, and rewriting. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11436–11458, Miami, Florida, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In Advances in Neural Information Processing Systems, volume 36, pages 46595–46623. Curran Associates, Inc.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. Preprint, arXiv:2310.17631.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Additional Information in Data Preparation

A.1 Crawling Sources

For Chinese part, we crawled data from the following high quality and reputable sources:

1. *Chinese Writer Website* (**CN Writer**, 中国作家网) [3] : this cite collects all publishable fictions, proses, poets from professional writers from China, powered by Chinese Association of Writer. The writings are all professionally written. The total number of raw data is approximately 5k.

2. *The pivot website for example essays* (**PW4ES**, 第一范文网) [4] : this cites collects numerous functional writing sources, such as contracts, plans, conclusions, thoughts, speeches and deliveries etc. The writings are of high quality and they serve as examples for learners. The total number of raw data is approximately 30k.

3. *September for example essays* (**SeptES**, 九月范文网) [5]: this cites complements to the above cites, with additional functional writings. The writings are of high quality and they serve as examples for learners. The total number of raw data is approximately 30k.

4. *Zhejiang Publicity* (**ZJPub,** 浙江宣传) [6] : this cites collects numerous argumentatives, critics targeting at social/historical/cultural affairs. These articles are targeting electronic self-media readers, and are written by professional newspaper writers. The total number of raw data is approximately 10k.

5. *Cite for Officials* (**Officials**, 公文网) [7]: this cites collects examples for official articles writings, including propaganda, deliveries, announcements, etc. We purchased the articles from the cite instead of crawling for its commercial use. The articles are written by expert civil servants from the government, and is of high quality. The total number of raw data is approximately 20k.

[3] https://www.chinawriter.com.cn/

[4] https://www.diyifanwen.com/

[5] https://www.chinesejy.com/

[6] https://zjnews.zjol.com.cn/zjxc/

[7] https://www.gongwen.com.cn/

For English part, we crawled data from the following high quality and reputable sources:

1. *American Rhetoric*[8] : This website records famous speeches in American history, including historical speeches as well as parliamentary speeches and questions.

2. *Obook*[9]: This website records numerous English published books with a wild range of genres, including fiction, prose, poet, novel across 16 century to contemporary.

3. *IvyPanda*[10]. This website serves top level example essays across 32 topics, including art, business, culture, environment, history, music and so on. We use huggingface datasetqwedsacf/ivypanda-essays [11] from the same source and the number is approximately 100K.

A.2 Included Writing Genres

Fiction : Fiction focuses on imaginative narratives, emphasizing character development, plot structure, and environmental depiction. It reflects social realities or human emotions, with a focus on details and conflicts driving the story forward.

Poetry : Poetry is characterized by line breaks, condensed language, and symbolic imagery, with an emphasis on rhythm and sound, as well as the intense concentration of emotion and thought.

Prose : Prose encompasses descriptive and imaginative writing without the constraints of poetic structure. It often explores themes and ideas in clear, expressive language, engaging the reader in a reflective or emotional experience.

Essay : A creative essay blends personal reflection and artistic style. It is often subjective, descriptive, and exploratory, focusing on an idea, experience, or insight in a unique and engaging way.

Argumentative : This writing builds a compelling case centered around a perspective or opinion, supported by logical reasoning or persuasive rhetoric. It seeks to convince the audience using passionate and effective arguments.

Report : A report is an objective, structured, and formal document that presents data, findings, and analysis of specific topics or activities, often following a standardized format.

Summary : Summarizing involves condensing large pieces of information into brief and concise overviews, focusing only on the key points, events, or ideas introduced in the original text.

Letter : A formal or informal written communication addressed to another person or entity, often following a clear structure that includes salutations, body content, and closing remarks.

Application : Applications are formal documents written in a specific format, expressing a request, often for employment, educational admissions, or permissions. They are brief and structured.

Speech : A speech is a prepared piece of writing meant to be spoken aloud, tailored for an audience, often persuasive or inspiring, and is structured to guide the listener through ideas or arguments.

Delivery : Delivery writing includes real-time or impromptu words, such as announcements or ceremonial addresses, meant for immediate and direct communication in specific events or contexts.

Plan : A plan outlines structured steps, timelines, or objectives to achieve a specific goal or outcome. It is often practical and formatted to organize resources and tasks effectively.

Contract : A contract is a formal, legal document outlining agreements between parties, specifying terms, responsibilities, and obligations, often in precise and enforceable language.

Official : Official writing refers to documents meant for administrative, governmental, or institutional purposes, often rigid in format and addressing formal matters or processes.

A.3 Leaf Node Traits Explained

We briefly introduce the leaf nodes traits in Table 7.

[8] https://www.americanrhetoric.com/top100speechesall.html

[9] https://www.obooko.com/

[10] https://ivypanda.com/

[11] https://huggingface.co/datasets/qwedsacf/ivypanda-essays

Traits	Description	Rubrics
Opening & Ending	Whether the opening and ending are engaging, with no abrupt stops or forced elevation of tone/plots/conclusions.	1-4: worse than reference 5-7: comparable to reference 8-10: Superior to reference
Language & Rhetoric	Using appropriate rhetoric, is the vocabulary and expression rich? Has a monotonous, list-like style been avoided?	
Proper instance	No violation to real-world knowledge. Whether proper instances are used to address argumentations.	
Argumentative & Logics	Whether the logic in arguments, plot development, and overall writing is appropriate and coherent. Ensure smooth transitions and avoid abrupt or forced causal connections.	
Emotion	Are the emotions effectively conveyed to the readers? Are the characters in the writing portrayed with appropriate emotional depth?	

Table 7: Illustration for different traits.

Traits	Description	Rubrics
Plots	Whether the plots are reasonable	1-4: worse than reference 5-7: comparable to reference 8-10: Superior to reference
Formatting	Checking all titles, lists in the writing with Regex. Detecting Chinese Titles, markdown titles, ordered lists, unordered lists.	0: Violation in hierachial relations, inproper unordered list in continuous texts 5. moderate titling or no titling are found 10: titling satisfies all the checks from the rules.
Paragraphing	Checking whether the paragraphs sectioning are reasonable or not.	1-4: Disproportionate paragraphing 5-7: paragraphing 8-10: paragraphing with superior deigns
Impression	Inspecting whether the writing satisfies the writing instructions theme and requirements.	1-4: worse than reference 5-7: comparable to reference 8-10: Superior to reference

Table 8: Illustration for Format and Impression traits.

B Further Discussions

B.1 Discussions on different evaluators

We further analyze the influence of Judge LLMs. We select the Level II and Level III tasks and compute the sample level Pearson correlation between GLM-4, Gemini-2.0-Flash, GPT-4o-1120, Deepseek-V3, Deepseek-R1. We concatenate all 3 inference model (GLM, Gemini and GPT) responses score as 3 times long vector and compute the Pearson correlation via it. Results are plotted in the form of heatmap in Figure 4. Results showed that Deekseek-V3 owns the highest pearson correlation with human judger, while GLM and GPT shares very poor correlation with human. On the other hand, LLM evaluators all showed very high correlation with each other ($\rho > 0.5$), indicating the common potential biases. Human experts reached $\kappa = 0.56$ and $\rho = 0.67$ in cross validation, confirming such gap between human and LLM Judges.

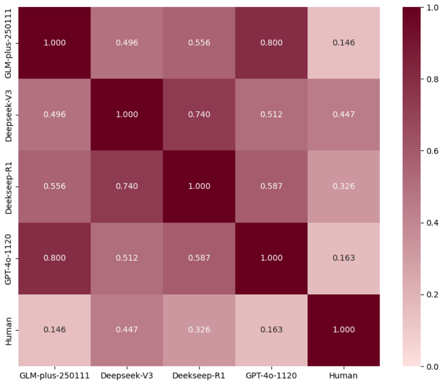

Figure 4: Pearson correlation cross evaluators and humans experts.

B.2 Discussion on Reference-based and Reference-free Evaluation

We experimented in a refined reference-free setting (by removing the existence of reference and re-judge) and compare it to the reference-based setting with a random and evenly picked subset from HowToBench (N=300). We calculated the system level correlation scores with all samples from 3 tasks altogether and summarize the results in Table 9.

From the experiment results, **ToW** still maintained high system level correlation while the base-

	Reference based			Reference Free		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Deepseek-baseline	0.688	0.717	0.556	0.592	0.383	0.278
Deepseek-rubric	0.585	0.533	0.389	0.452	0.333	0.222
Deepseek-CoW	**0.748**	**0.750**	0.556	**0.714**	**0.733**	**0.556**
Gemini-baseline	0.160	0.267	0.222	0.697	0.750	0.556
Gemini-rubric	0.716	0.733	0.556	0.423	0.600	0.389
Gemini-CoW	**0.723**	**0.767**	0.556	**0.749**	**0.800**	**0.611**

Table 9: Reference-based evaluation and Reference-free evaluation results.

	Human	**Deepseek-V3**	**GPT-o3-mini**	**Claude-3.5**
Deepseek-baseline	0.688	0.477	0.723	0.756
Deepseek-rubric	0.585	0.451	0.652	0.646
Deepseek-CoW	**0.748**	**0.607**	**0.816**	**0.783**
Gemini-baseline	0.160	**0.730**	0.459	0.380
Gemini-rubric	0.716	0.580	0.469	0.473
Gemini-CoW	**0.723**	0.715	**0.528**	**0.552**

Table 10: Influence on system level correlation from reference sources.

line, rubric methods drops with the absence of reference. This indicates that chain-of-writing can judge without reference, which goes beyond the rubric scoring methods.

B.3 Discussion on Reference Source

One of the core principles of HOWTOBENCH is the reliance on high-quality human experts and writers as references for evaluation. We investigate the feasibility and reliability of using LLM-generated texts as references and assess their credibility at the system level.

Specifically, we adopt a setting where the instructions and guiding information in HOWTOBENCH remain unchanged, but the inference output of a particular LLM is used as a 6 point reference to guide evaluation. We employ Gemini-2.0-Flash as the Evaluator and compare the results against human references as well as those generated by Deepseek-V3, GPT-4-o3-mini, and Claude-3.5-sonnet-1022. The three models are recognized for their strong performance in writing tasks. The system-level correlations are summarized in Table 10. References derived from alternative sources generally result in lower consistency rates, whereas human references achieve significantly higher agreement. Furthermore, the **ToW** demonstrates robustness across references of varying origins, indicating that its effectiveness is independent of the reference source.

C Full Plots for Analysis sections

C.1 Plots Between Input Length, Output Length and Scores

Figure 5 presents the scatter and linear regression between input length, output length, overall scores and content scores.

C.2 Edge Weights across Multiple Genres

D Prompts for writing genre classifier

Classifier Prompt

Input: text

Please classify the following written text. You will be provided with a text. Based on the given framework, please categorize the text into one of the following categories:

- Creative - Fiction : Fiction focuses on imaginative narratives, emphasizing character development, plot structure, and environmental depiction. It reflects social realities or human emotions, with a focus on details and conflicts driving the story forward.

- Creative - Poetry : Poetry is characterized by line breaks, condensed language, and symbolic imagery, with an emphasis on rhythm and sound, as well as the intense concentration of emotion and thought.

- Creative - Prose : Prose encompasses descriptive and imaginative writing without the constraints of poetic structure. It often explores themes and ideas in clear, expressive language, engaging the reader in a reflective or emotional experience.

- Creative - Essay : A creative essay blends personal reflection and artistic style. It is often subjective, descriptive, and exploratory, focusing on an idea, experience, or insight in a unique and engaging way.

- Creative - Argumentative : This writing builds a compelling case centered around a perspective or opinion, supported by logical reasoning or persuasive rhetoric. It seeks to convince the audience using passionate and

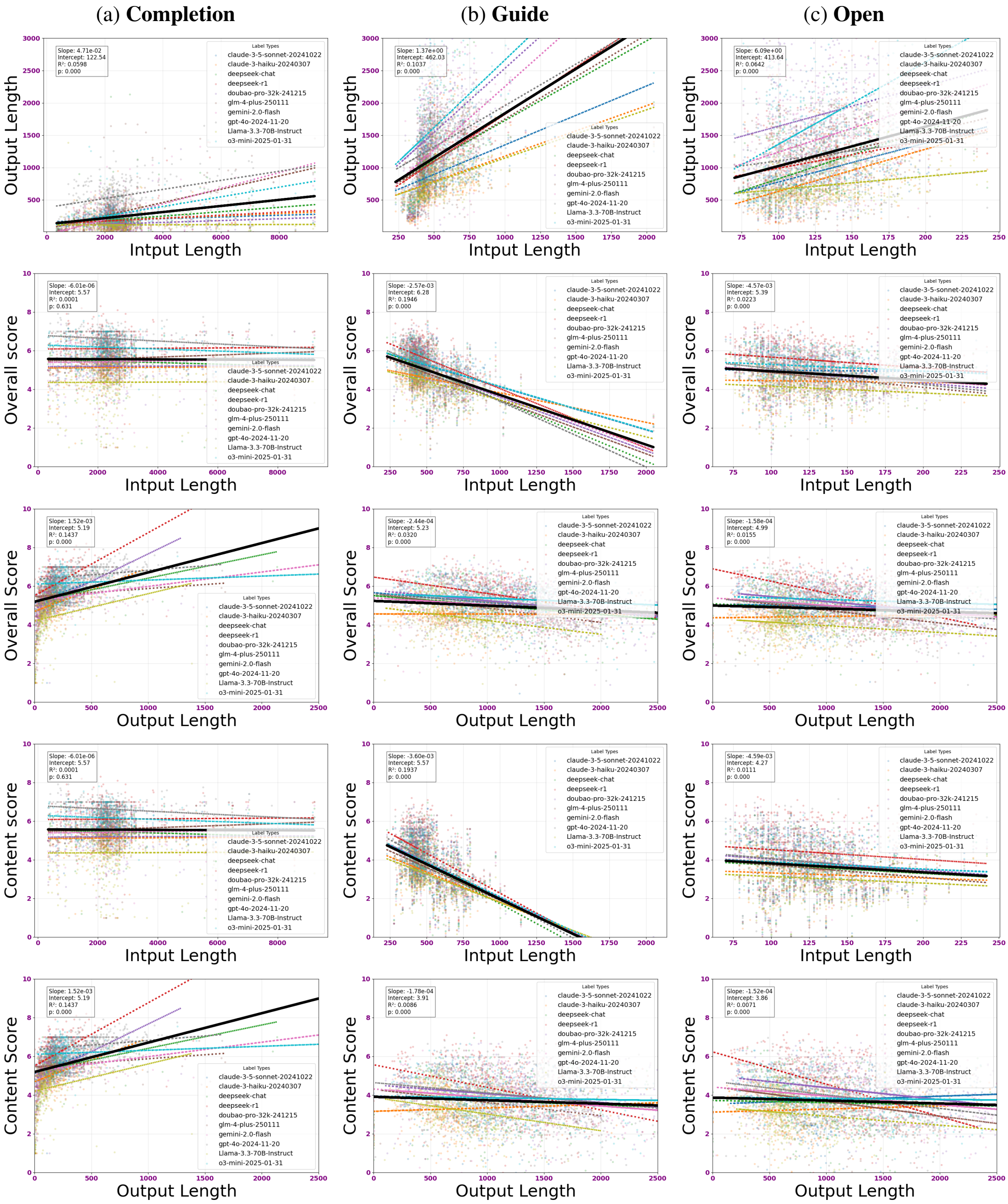

Figure 5: Factor Analysis between input length, output length, overall score, content score. The bold black line indicates the regression results from all LLM data points.

effective arguments.

- Functional - Report : A report is an objective, structured, and formal document that presents data, findings, and analysis of specific topics or activities, often following a standardized format.

- Functional - Summary : Summarizing involves condensing large pieces of information into brief and concise overviews, focusing only on the key points, events, or ideas introduced in the original text.

- Functional - Letter : A formal or informal written communication addressed to another person or entity, often following a clear structure that includes salutations,

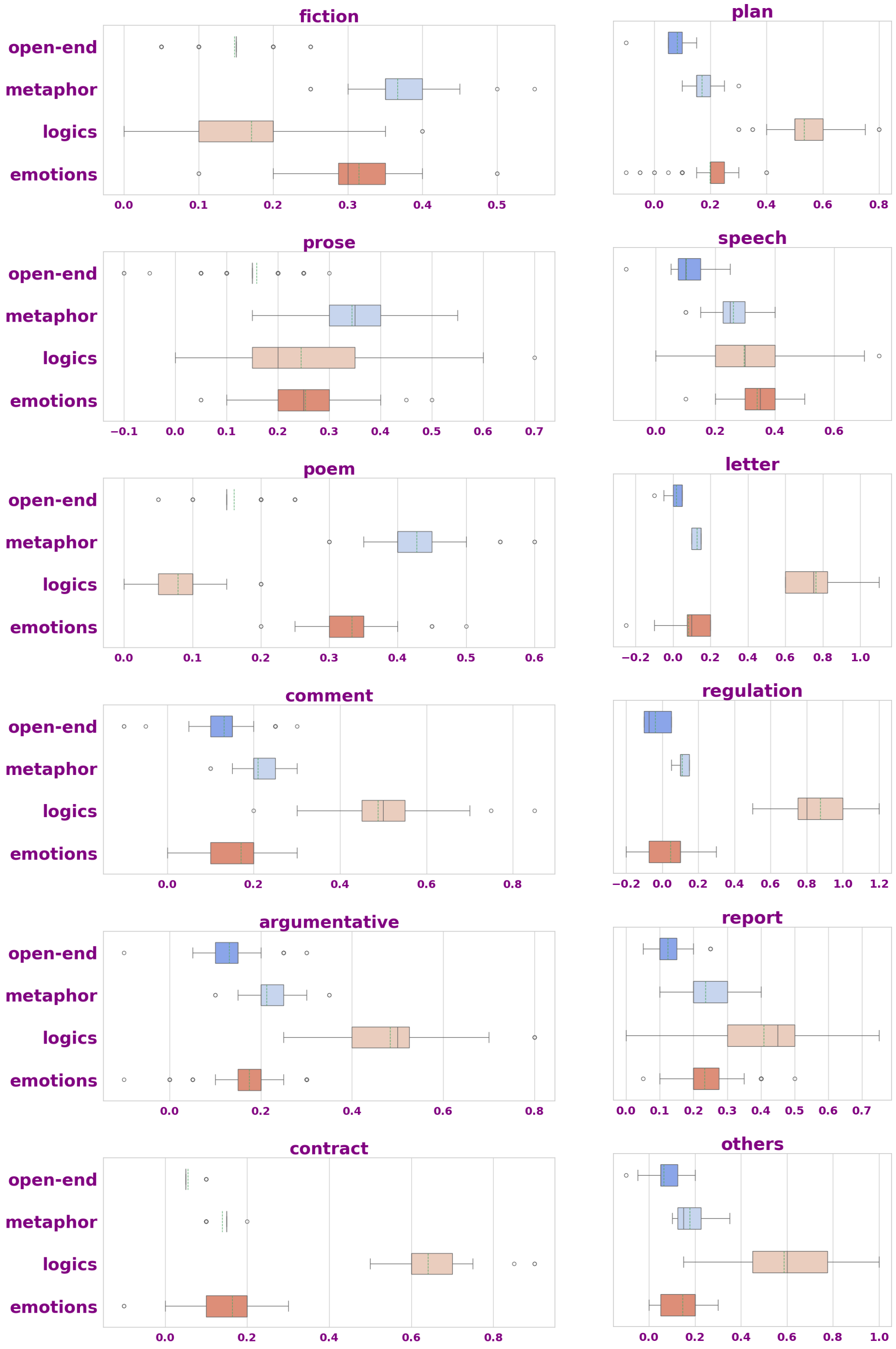

Figure 6: Edge weight distribution on different genres. The wider is the box horizontally, the more varied is the corresponding weight within the genre.

body content, and closing remarks.

- Functional - Application : Applications are formal documents written in a specific format, expressing a request, often for employment, educational admissions, or permissions. They are brief and structured.

- Functional - Speech : A speech is a prepared piece of writing meant to be spoken aloud, tailored for an audience, often persuasive or inspiring, and is structured to guide the listener through ideas or arguments.

- Functional - Delivery : Delivery writing includes real-time or impromptu words, such as announcements or ceremonial addresses, meant for immediate and direct communication in specific events or contexts.

- Functional - Plan : A plan outlines structured steps, timelines, or objectives to achieve a specific goal or outcome. It is often practical and formatted to organize resources and tasks effectively.

- Functional - Contract : A contract is a formal, legal document outlining agreements between parties, specifying terms, responsibilities, and obligations, often in precise and enforceable language.

- Functional - Official : Official writing refers to documents meant for administrative, governmental, or institutional purposes, often rigid in format and addressing formal matters or processes.

The following is the text to be classified:

{text}

Present your judgment in double bracket enclosed form, such as [[Creative - Fiction]].

E Prompts for Coarse Rubric Scoring Filter

Scoring Filter Prompt

Input: content

Please act as a professional fiction reviewer to evaluate the following novel and rate it based on the specified dimensions. For each dimension, assign a score between 1 and 5 and provide a brief explanation. Finally, give the fiction a total score (between 1 and 5).

[Fiction Start]

{content}

[Fiction End]

[Criteria Start]

1. Plot and Structure

- Compactness of the Plot : Are the plotlines smooth and tight? Do they hold enough allure to sustain the reader's interest?
- Structural Layout : Is the novel's structure reasonable? Does it avoid excessive drag or hollow portions in the story? For medium to long-form novels, are there clear stages of exposition, rising action, climax, resolution, and reversals?
- Sense of Rhythm : Is the progression of the story balanced? Does the unfolding of events carry tension and momentum, especially in medium to long-form novels, where pacing is critical?

2. Character Development

- Depth of Characters : Are the characters well-rounded and multi-dimensional? Do they exhibit unique personalities and undergo meaningful changes?
- Character Growth : Does the novel reasonably portray the growth, transformation, or conflicts of its characters? Are there evident internal struggles or character arcs?
- Character Relationships : Are the interactions between characters natural? Do they contribute meaningfully to the advancement of the plot?

3. Themes and Ideas

- Depth of Theme : Does the novel have a clear and compelling theme? Is the theme substantial and thought-provoking?
- Expression of Ideas : Does the novel

convey profound thoughts or ideas through its characters, plot, or symbolic elements? Does it inspire reflection in its readers?
- Social and Cultural Context : Does the novel provide deep insight into a particular era, society, or culture through its story and characters?

4. Language and Writing Style

- Language Style : Is the author's language vivid and elegant? Can it effectively convey the emotions and thoughts of the characters?
- Adaptability of Language : Does the language align with the story's atmosphere and context? Does it enhance the emotional intensity of the novel?
- Detail Description : Are the descriptive details fitting and appropriate? Do they aid in character-building, setting the mood, or driving the story forward?

5. Emotional Resonance

- Emotional Depth : Does the novel evoke emotional resonance in readers? Can it make readers empathize and emotionally invest in the story?
- Authenticity of Emotions : Are the emotions in the novel realistic and believable? Do they have the power to move the reader?

6. Innovation and Uniqueness

- Innovative Elements : Does the novel showcase originality in some areas? Does it challenge traditional narrative conventions or stylistic norms?
- Unique Perspective : Does it approach a topic or tell its story from a distinctive angle? Does it reflect a strong, memorable voice or personality?

[Criteria End]

Begin your evaluation by assigning a score between 1 and 10 for each dimension, along with a brief explanation. Conclude with the novel's overall score (1 to 5). A score

	CN Writer	PW4ES	SeptES	ZJPub	Officials
1	0	153	20	0	2
2	12	351	134	3	15
3	1137	13188	10261	272	8705
4	4468	72908	3216	722	6957
5	19	861	73	86	204

Table 11: Filter score from the coarse rubric scoring system implemented with Claude-3-5-sonnet-1022.

of 1–2 indicates the dimension performed poorly, 3-4 means it was average, and 5 means it excelled in the dimension. Please use the following example output format:

"Plot and Structure": 2
"Character Development": 3
"Themes and Ideas": 4
"Language and Writing Style": 3
"Emotional Resonance": 3
"Innovation and Uniqueness": 2
"Overall Rating": 3

Table 11 lists the score distribution from the filter.

F Prompts for Back-Construction

Back Construction Prompt

Input: content

Assume that you are to provide instructions to a large language model, asking it to generate the following fiction. Provide detailed instructions with the following structure:

1. Plot and Structure: Summarize the main content of the fiction in one sentence of no more than 100 words.
2. Character Development: Describe the personalities, experiences, and relationships of the main characters in no more than 100 words per character, with a maximum of 5 characters in total.
3. Theme and Message: Summarize the theme and message the fiction aims to convey in no more than 100 words.
4. Language and Style: Describe the overall

linguistic style of the fiction and the level of detail in its descriptions, in no more than 100 words.
5. Emotional Resonance: Specify the type of emotional resonance the fiction aims to evoke in readers in no more than 100 words.
6. Innovation and Originality: Describe how the fiction should demonstrate uniqueness or originality in no more than 100 words.

Output the instructions using the following format:

<Plot and Structure Start>
xxxx
<Plot and Structure End>

<Character Development Start>
xxxx
<Character Development End>

<Theme and Message Start>
xxxx
<Theme and Message End>

<Language and Style Start>
xxxx
<Language and Style End>

<Emotional Resonance Start>
xxxx
<Emotional Resonance End>

<Innovation and Originality Start>
xxxx
<Innovation and Originality End>

Please base your response on the following target fiction.

[Target Fiction Start]

{content}

[Target Fiction End]

G Human Picking Guideline

G.1 Task Description

Your task is to evaluate and compare four different writings based on a provided writing instruction. Each writing is a response to the same instruction, and your goal is to pick the one that fits the instruction with the highest quality. Use the evaluation criteria provided below to make your judgment. The selected writing should be the one that most effectively fulfills the writing instruction and demonstrates the highest level of quality across both content and format.

G.2 Annotation Fields

G.2.1 Visible Inputs

- **Writing Instruction** : A clear description of the requirements or objectives for the writing task (e.g., structure, tone, purpose, or audience).

- **Guiding Information** : If applicable, specific details that the writings are expected to follow (e.g., key points, required examples, or constraints). For tasks requiring "guide generation," ensure the writings strictly adhere to these details.

- **Writing 1/2/3/4** : The individual LLM writings submitted for judging.

G.2.2 Your Observations

- Write down notes on how each writing satisfies the instruction and aligns with the evaluation criteria.

- Highlight specific strengths and weaknesses of each writing that influenced your judgment.

G.2.3 Annotation Process

Step 1: Read Each Writing Thoroughly

- Carefully read each writing submission. - Pay attention to how well the author has addressed the writing instruction and incorporated the guiding information provided. - Consider the quality of the arguments, organization, and style of each piece. Make sure to read thoroughly before forming a judgment.

Step 2: Apply the Quality Criteria

- Systematically assess each writing response against the evaluation criteria outlined below. - Use both content and format criteria to conduct your evaluation and determine the strengths and weaknesses of each submission. - You may apply a pointwise scoring system (e.g., rating each category from 1 to 5) to help you compare the writings more quantitatively. These scores should support — but not replace — your final judgment.

Step 3: Select the Best Writing

- Based on your evaluation in Step 2, determine which writing best fulfills the writing instruction and meets the specified quality criteria. - Document your reasoning for selecting the chosen writing. Highlight why the selected piece was superior and what weaknesses were present in the others.

G.3 Evaluation Criteria

Your evaluation should be based on two main areas: Content and Format . Each area contains specific criteria to guide your assessment:

G.3.1 Content

1. Theme/Argument/Topic Fit :

- How well does the writing address the objective of the instructions?

- Are the arguments or ideas relevant and clearly aligned with the given topic?

- Does the writing stay focused, or does it go off-topic?

2. Tone and Language :

- Is the tone appropriate for the audience and purpose outlined in the writing instruction?

- Does the writing use clear, engaging, and professional language where required?

- Is the tone consistent throughout the piece?

3. Attractiveness of Opening and Profound Ending :

- Does the writing start with a strong and engaging opening that catches the reader's attention?

- Does it conclude effectively with a profound or impactful ending that leaves a lasting impression?

4. Rhetoric, Logic, and Examples :

- Does the writing employ effective rhetoric (e.g., persuasive techniques, vivid imagery, or strong analogies)?

- Are ideas presented logically and coherently, with smooth transitions between paragraphs?

- Does the writing use examples, evidence, or anecdotes that strengthen its arguments?

G.3.2 Format

1. Basic Format Requirements of the Genre

- Does the writing follow the structural conventions of the specified genre (e.g., essay, article, guide, etc.)?

- Are any mandatory elements of the format (e.g., headings, bullet points, or lists) included and used appropriately?

- Avoiding Abrupt Bullets or Unordered Lists :

- Does the writing avoid disorganized or improperly formatted lists or bullet points that disrupt the flow of the content?

- Are lists used sparingly and only when they enhance clarity?

2. Adequate Titling and Subtitle Structures

- Does the writing include an appropriate, engaging, and informative title?

- If subtitles are required or used, are they logical, helpful, and aligned with the overall structure of the piece?

G.3.3 Additional Considerations

- **Consistency with Instruction and Guiding Information**

Always double-check whether the writing adheres to the writing instruction and any specific guiding information provided. A failure to follow core requirements should result in a lower ranking.

- **Avoid Personal Bias**

Focus on the objective quality of the writing, not on personal preferences or subjective interpretations that are unrelated to the task.

- **Use a Systematic Approach**

Ensure that you assess each writing fairly and systematically using the outlined evaluation criteria. If you're unsure between two submissions, revisit the instruction and criteria to resolve ambiguity.

H Rubric Prompts for LLM-based Evaluation

H.1 Argumentative

1. Clarity of the Theme and Argument

Clarity of the Theme : Is the theme of the essay clear and prominent? Can readers quickly grasp the central idea? Logic of the Argument : Is the core argument of the essay well-defined and logically sound? Does it effectively support the overall content?

2. Adequacy and Diversity of Evidence

Adequacy of Evidence : Does the essay provide enough persuasive evidence? Is the evidence specific, detailed, and closely related to the theme? Diversity of Evidence : Are the types of evidence varied (e.g., theoretical analysis, factual examples, data citations, expert opinions)? Does the evidence approach the theme from multiple perspectives?

3. Language and Logical Expression

Language Expression : Is the language of the essay concise, clear, and logical? Are the sentences coherent and easy to understand? Does the language enhance the essay's persuasiveness? Clarity of Logic : Is the reasoning process rigorous and progressive, leading to strong and rational arguments?

4. Structure and Writing Logic

Structural Coherence : Is the structure of the essay clear and well-organized? Does it follow a logical format, such as "introduction-body-conclusion" or parallel argumentation? Consistency in Flow : Are the paragraphs cohesive and logically arranged? Does the essay use effective transitions to strengthen the cohesiveness and persuasiveness of its arguments?

5. Reflectiveness and Innovation

Depth of Reflection : Does the essay demonstrate some degree of reflection on societal, individual, or universally relevant issues? Does it inspire deeper thinking in readers? Novelty of Perspective : Are the arguments innovative or distinctive? Does the essay present surprising or original viewpoints or methods of argumentation?

H.2 Summary

1. Goals and Depth of Reflection

Clarity of Goals : Does the summary clearly articulate the specific objectives and plans of the work? Does it effectively review and analyze according to the established goals? Depth of Reflection : Does the summary deeply reflect on the achievement of the goals? Does it extract meaningful lessons from successes or shortcomings to guide future actions?

2. Content and Logic

Comprehensiveness of Content : Does the summary cover the key aspects of the work process? Does it address important outcomes, challenges, and areas for improvement in detail? Clarity of Logic : Is the content presented in a well-structured and logical manner? Is it organized by criteria such as timeline, importance, or category? Is it easy for readers to follow and capture the key points?

3. Language and Precision

Conciseness of Expression : Is the summary written with precise and concise language? Is it effective in conveying information within a limited space? Persuasiveness of Language : Does the language inspire trust and resonance? Is it engaging and persuasive enough to capture the reader's attention?

4. Structure and Readability

Rationality of Structure : Is the structure of the summary clear and reasonable (e.g., having clear headings and well-distributed paragraphs)? Does it enhance the overall reading experience? Aesthetic Presentation : Does the summary use visual elements like clear formatting, highlighted keywords, or data references to improve the effectiveness of information delivery?

5. Innovation in the Summary

Uniqueness of Analytical Perspective : Does the summary demonstrate the author's unique insights or thought-provoking analysis? Does it break away from traditional formats to showcase individual or team creativity? Foresight in Recommendations : Does the summary propose specific and forward-thinking suggestions or future plans? Does it combine past experiences and trends to provide meaningful guidance?

H.3 Contract

1. Integrity and Clarity

Clause Coverage : Do the contract provisions comprehensively address all necessary aspects, including the rights and obligations of both parties, liability for breach, and dispute resolution mechanisms? Have important details been thoroughly included to avoid omissions? Language Clarity : Is the contract language concise and clear? Does it avoid ambiguity and multiple interpretations, ensuring both parties can accurately understand its terms?

2. Legality and Risk Control

Legal Compliance : Does the contract fully comply with relevant laws and regulations, including those related to the qualification of parties, jurisdiction, and compensation mechanisms? Has the contract considered specific legal requirements in its respective field, such as labor laws or intellectual property laws? Risk Prevention : Does the contract effectively mitigate potential legal loopholes or risks of breach? Are its terms designed with a thorough assessment of legal risks and reasonable strategies for their avoidance?

3. Practical Operability

Execution Details : Does the contract provide detailed considerations for implementation, covering specific aspects like payment methods, delivery standards, and service quality? Does it offer clear operational guidelines and responsibilities for the performance process? Performance Monitoring : Does the contract include provisions for monitoring implementation, facilitating both parties to manage and fulfill their respective obligations effectively?

4. Balance and Fairness

Equity Balance : Does the contract reasonably balance the rights and interests of both parties? Does it avoid obviously one-sided terms, such as unfair allocations of liability for breach or overly stringent conditions? Fairness of Design : Are the contract terms structured to reflect fairness and impartiality, effectively reducing the likelihood of disputes or conflicts?

5. Future Adaptability and Sustainability

Flexibility for Adjustment : Does the contract account for potential future changes in circumstances, such as legal amendments or market fluctuations? Does it offer flexible provisions for modifications or adjustments to address unforeseen developments? Long-Term Cooperation Potential : Does the contract safeguard the potential for long-term collaboration? Are the terms designed with sustainability in mind, avoiding rigidity that might hinder future partnerships?

H.4 Delivery

1. Linguistic Expression

Clarity of Expression : Is the speech language clear, concise, devoid of redundancy, and easy to understand? Are grammar and syntax correct, with varied and layered sentence structures? Appropriateness of Language : Does the expression align with the demands of the occasion, employing a formal, humorous, or emotional style as needed for the specific context?

2. Emotional Expression and Impact

Sincerity of Emotion : Does the speech convey authentic and profound emotions, reflecting the speaker's genuine attitude? Emotional Resonance : Does the content resonate with the audience, evoke emotional engagement, and fit the tone of different occasions?

3. Logical Structure and Coherence

Structural Clarity : Is the speech well-structured, with a clear introduction, body, and conclusion? Are key points highlighted, and does the flow of ideas remain coherent? Natural Transitions : Are the transitions between sections logical and smooth, ensuring content flows naturally?

4. Suitability for the Occasion

Relevance of Content : Does the speech align with the specific theme and atmosphere of the occasion (e.g., weddings, memorials)? Audience Consideration : Does the speech take into account the audience's psychology and needs, with language and expression respectful of the context and culture?

5. Creativity and Originality

Unique Perspective : Does the speech reflect the speaker's creativity or unique perspective, rather than relying entirely on conventional templates? Memorable Impressions : Are there innovative expressions or distinctive personal elements that leave a lasting impression and highlight the speech's individuality?

H.5 Documentary

1. Authenticity and Factual Accuracy

Does the work accurately and faithfully reflect historical events or social phenomena, based on thorough investigation and research with reliable sources? Does the work present the complexity of events from multiple perspectives, avoiding bias while maintaining factual rigor?

2. Characterization and Emotional Expression

Are the characters multidimensional and well-developed, reflecting their inner world and emotional changes convincingly? Are the relationships between characters intricate and dynamic, contributing to story development, and are the characters' growth or transformations reasonable and compelling?

3. Structure and Narrative Techniques

Is the overall narrative structure clear and logical? Are the plot and pacing engaging and well-balanced, avoiding excessive length or repetitiveness? Does the work effectively use techniques such as nonlinear timelines, spatial transitions, or shifts in perspective and detail to enhance storytelling and literary quality?

4. Ideological Depth and Social Significance

Does the work encourage readers to deeply reflect on social phenomena, historical contexts, or human behaviors, demonstrating a strong sense of social concern? Does it display critical and reflective perspectives, courageously exposing social issues and engaging in an in-depth exploration of history or society?

5. Language and Writing Style

Is the language concise, clear, and expressive, employing techniques such as detail, metaphor, or description to enhance literary quality and emotional impact? Does the narrative style align with the theme and emotions of the work, enhancing its readability and artistic value?

H.6 Essay

1. Argument and Depth of Thought

Core Argument : Does the review article present a clear and well-defined central argument or position? Does it effectively and directly address the topic or text in question? Depth of Thought : Does the article demonstrate profound insight into the subject or material? Does it employ thorough analysis or critical thinking to deliver meaningful viewpoints?

2. Logic and Evidence

Clarity of Logic : Is the argument logically coherent? Is the article well-structured and organized, unfolding its analysis in a systematic and layered manner? Quality of Evidence : Does the article provide strong evidence to support its central argument? Is the evidence thoroughly analyzed and interpreted in a persuasive way?

3. Language and Style

Language Precision : Is the language used accurate, concise, and persuasive? Does it reflect the analytical nature of commentary writing? Distinctive Style : Does the writing style demonstrate critical thinking? Does it reflect the author's depth of thought and an individualized approach to expression?

4. Perspective and Comprehensiveness

Multifaceted Analysis : Does the article analyze and interpret the topic or text from multiple perspectives, reflecting a comprehensive understanding of the issue? Comprehensiveness : Does the review integrate various layers of analysis, presenting a holistic grasp of the subject matter?

5. Originality and Thought-Provocation

Originality : Does the article present unique insights or novel perspectives? Does it offer new ways of thinking or intellectual contributions to the discussion? Thought-Provocation : Does the content of the review inspire further reflection or exploration by the reader? Does it open up new interpretative possibilities for the topic under discussion?

H.7 Fiction

1. Plot and Structure

Plot Coherence : Is the plot well-paced and engaging? Does it maintain the reader's interest? Structural Design : Is the structure of the novel logical? Are there instances of unnecessary delays or plot gaps? For medium- to long-length novels, a clear progression (beginning, development, turning points, climax, and resolution) is crucial. Rhythm and Balance : Is the story progression well-balanced? Does the unfolding of events create narrative tension? Proper pacing is especially critical for medium- and long-length works.

2. Characterization

Character Depth : Are the characters well-developed, multidimensional, and distinct in personality? Character Development : Do the characters undergo meaningful growth, change, or conflict in a well-reasoned way? Are there clear internal struggles or character arcs? Interpersonal Dynamics : Are the interactions between characters natural? Do these relationships effectively drive the plot forward?

3. Themes and Ideas

Thematic Depth : Does the novel have a clear theme? Is the theme explored with sufficient depth and intellectual value? Ideological Expression : Does the novel convey profound ideas through characters, plot, or symbols? Does it provoke critical thought? Social and Cultural Context : Does the story reflect a nuanced understanding of a particular era, society, or culture through its narrative and characters?

4. Language and Prose

Style of Expression : Is the author's language vivid, elegant, and effective in portraying the emotions and thoughts of the characters? Contextual Adaptation : Does the language align with the tone and atmosphere of the story? Does it enhance the emotional tension? Detailing : Are the descriptions appropriate and well-crafted, contributing to characterization, atmosphere, or plot progression?

5. Emotional Resonance

Emotional Impact : Does the novel evoke emotional resonance in readers? Does it foster empathy and emotional engagement? Emotional Authenticity : Are the emotions in the story realistic and compelling? Do they effectively move the reader?

6. Innovation and Distinctiveness

Originality : Does the novel exhibit creativity or innovation by breaking away from conventional tropes or styles? Unique Perspective : Does the novel present a distinct viewpoint or approach to exploring its subject matter? Does it convey a strong sense of identity and uniqueness?

H.8 Letters

1. Structure and Format

Does the letter follow standard formatting with appropriate salutation, body, and closing? Is the letter's structure clear, with distinct paragraphs and a logical flow? Is the letter well-organized and visually appealing, making it easy to read?

2. Language Brevity and Clarity

Is the language in the letter concise, avoiding long and complex sentences? Is the expression clear, is the logic coherent, and is the information accurate? Are ambiguities and unclear statements avoided to ensure the recipient's full understanding?

3. Tone and Attitude

Is the tone appropriately chosen based on the recipient's identity and the letter's purpose? Does the tone convey sincerity and respect? Does the letter maintain the necessary politeness and professionalism?

4. Clear Purpose and Accurate Content

Is the core purpose of the letter (e.g., request, notification, suggestion) clearly expressed? Is the content accurate and free from errors or ambiguous expressions? Does the letter stay focused on its goal without deviating from its theme?

5. Etiquette and Adaptability

Does the letter adhere to basic etiquette norms? Is the language and expression appropriate for the cultural context or situational needs? Is the overall visual presentation of the letter tidy, standardized, and easy to read?

H.9 Officials

1. Accuracy and Completeness of Content

Is the content of the document factual and accurate? Does it include all necessary information and details? Is there assurance that no critical parts are omitted? Does it comply with current laws, policies, and regulations?

2. Structure and Logical Flow

Is the structure of the document clear and reasonable? Is there a good logical connection between paragraphs? Is the sequence of information arranged logically? Does the content flow naturally without redundancy or confusion?

3. Language Standardization and Conciseness

Does the language conform to formal document standards? Are colloquial expressions avoided? Is the expression precise and rigorous? Is the language concise and clear, facilitating reader understanding and execution?

4. Formatting and Formality

Does the document follow standard formatting? Are sections like type, title, number, date, and signatory in compliance with requirements? Is the layout orderly, with correct punctuation and wording? Is the overall tone of the document formal and appropriate?

5. Executability and Legal Compliance

Does the document have clear executable directives? Are the proposed requirements and measures specific and actionable? Does the content comply with laws and regulations? Is there an assurance that it avoids any violations of law or public interest?

H.10 Plan

1. Clarity of Objectives

Core Objectives: Does the plan have clearly defined goals? Are the objectives measurable and achievable, effectively guiding execution? Detailed Objectives: Does the plan outline problem-specific solutions with well-defined, quantifiable indicators (e.g., percentage of sales growth, training completion rate)?

2. Feasibility and Executability

Execution Details: Does the plan provide clear operational guidance and a complete implementation process? Are specific implementation steps, timelines, and responsibilities clearly outlined? Execution Support: Does the plan account for key factors such as resources, personnel, and time during execution? Does it include contingency plans to address challenges?

3. Innovation and Differentiation

Unique Perspective: Does the plan break conventional approaches, offering fresh perspectives or solutions? Does it incorporate novel ideas, methods, or technological support? Innovative Value: Compared to existing plans, does the new plan demonstrate differentiation, effectively addressing issues or offering breakthrough solutions?

4. Risk Assessment and Mitigation Measures

Risk Identification: Does the plan identify potential risks and scenarios that could impact implementation? Mitigation Strategies: Does the plan propose concrete measures or alternative strategies to manage identified risks? Does it account for adaptability in addressing different scenarios?

5. Effectiveness Evaluation and Feedback Mechanism

Evaluation Tools: Does the plan include a comprehensive assessment mechanism to monitor outcomes, provide regular feedback, or track results over time? Optimization Capability: Does the plan incorporate mechanisms for adjustment and iteration based on practical feedback to ensure continuous improvement during implementation?

H.11 Poem

1. Language and Expressiveness

Innovation and Simplicity: Modern poetry often emphasizes linguistic innovation and unique expressiveness. When evaluating, focus on whether the poem uses distinctive language and effectively conveys rich emotions or ideas succinctly. Rhythm and Sound: Even without traditional rhymes, modern poetry enhances expression through rhythm and intonation. Evaluation should consider the flow of the poem's rhythm, the harmony of its sounds, and how these elements enhance emotional expression.

2. Theme and Depth of Thought

Philosophical and Reflective Qualities: Modern poems often explore profound themes such as individuality, society, and existence. Evaluation should assess whether the poem possesses philosophical or reflective qualities and whether it provokes thought in the reader. Uniqueness of Theme and Presentation: Attention should be given to whether the poem offers a unique perspective on its theme and employs metaphors or symbols rather than straightforward statements.

3. Emotional Expression and Nuance

Sincerity and Complexity of Emotion: Modern poetry typically conveys emotions indirectly, using nuanced language, symbolism, and implications. Evaluation should consider the sincerity of the emotions and whether the emotions exhibit complexity or depth. Integration of Emotion and Theme: Consider whether the emotional expression is tightly linked to the theme and whether the fluctuations and internal conflicts of the emotions enhance the poem's expressive power and depth of thought.

4. Uniqueness of Form and Structure

Innovative and Organic Structure: Modern poetry often features diverse structures, including fragmented or non-linear forms. Evaluation should note whether the poem's structure is innovative and effectively supports its theme and emotional expression. Unity of Form and Content: Modern poetry's form typically complements its content. Evaluation should consider whether the form strengthens the poem's inherent meaning and whether unique structures and layouts enhance expressive effect.

5. Overall Effect and Ambiguity

Artistic Effect and Interpretative Space: Modern poetry often has openness and ambiguity. Evaluation should consider the poem's overall effect—whether it resonates emotionally with the reader and stimulates diverse interpretations and reflections. Impact and Intellectual Provocation: Ultimately, the evaluation of a modern poem should consider whether it leaves a lasting impression on the reader, either through emotional impact or intellectual challenge.

H.12 Prose

1. Theme and Depth of Thought

Core Idea : Does the essay present a clear theme or central idea? Does it provoke readers to think deeply? Depth of Thought : Does the essay explore profound philosophical, social, or life-related issues? Does it use detailed descriptions or personal experiences to convey broader reflections?

2. Language and Style

Expression : Is the language concise, elegant, and expressive? Does it align with the characteristics of an essay, demonstrating literary quality and fluency? Unique Style : Does the writing exhibit a distinctive style or personal touch? Does it employ rhetorical techniques to convey the author's unique perspectives or artistic sensibilities?

3. Structure and Rhythm

Structural Coherence : Is the structure of the essay clear and well-organized? Does it effectively support the development of the theme? Sense of Rhythm : Is the pacing appropriate with a balanced flow? Does the arrangement of paragraphs and sentence structures enhance the reading experience?

4. Emotion and Impact

Authenticity of Emotion : Are the emotions in the essay genuine and profound? Does it move the reader through nuanced descriptions and emotional transitions? Emotional Resonance : Do the emotions in the essay resonate with readers? Does it possess universality or the power to emotionally engage its audience?

5. Cultural Context and Innovation

Cultural Depth : Does the essay reflect the author's understanding and contemplation of specific cultural, social, or historical contexts? Does it capture the spirit of the times or convey humanistic concerns? Innovation : Are the perspectives or expressions in the essay distinctive? Does it provide readers with new ways of thinking or unique literary experiences?

H.13 Report

1. Structure and Logical Coherence

Clarity of Structure: Is the report's structure clear? Are the contents organized in a hierarchical and logical manner? Does the sequence guide the reader toward a step-by-step understanding? Content Coherence and Logic: Are the sections well-connected? Does the report avoid issues of repetition or omission? Is the overall logic rigorous, and is the narrative smooth and consistent?

2. Accuracy and Completeness of Content

Information Accuracy: Are the data and information in the report accurate, reliable, and based on credible sources? Do they align with objective facts, without contradictions or errors? Content Completeness: Does the report cover the core aspects of the topic and provide comprehensive background information? Are any key points omitted?

3. Language and Writing Quality

Precision and Conciseness: Is the language clear and concise, avoiding unnecessary verbosity? Are grammar and spelling correct? Formality and Style: Does the writing adhere to formal academic standards? Is the expression professional and fluent?

4. Innovation and Depth

Innovation: Does the report offer fresh perspectives, insights, or methods? Does it demonstrate creativity by providing a novel approach or new angle to the problem? Depth of Content: Does the report delve into the essence of the problems rather than staying at a superficial level? Does it reflect high analytical capability and research depth?

5. Relevance and Practicality

Alignment with the Theme: Does the content closely align with the report's theme? Does it address the purpose of the report and meet the needs of the intended audience? Practical Value: Are the suggestions or conclusions actionable? Can they provide meaningful help or references for the target audience?

H.14 Document

1. Structural Integrity and Organization

Structural Standards : Does the document follow a complete and standard format (e.g., title, background, main body, conclusion)? Is it well-organized and logically coherent? Are the transitions between paragraphs smooth? Logical Organization : Is the content arranged in a reasonable manner to facilitate quick understanding and response from the reader? Does it comply with conventional document writing standards?

2. Conciseness and Clarity of Expression

Accuracy of Expression : Is the language concise and the information clearly conveyed? Are the word choices accurate? Does the document avoid overly long, complex sentences or ambiguous statements? Effective Communication : Does the document achieve the goal of delivering information quickly and clearly, while minimizing unnecessary ambiguity and the need for revisions?

3. Norm Compliance and Formatting Consistency

Format Compliance : Does the document strictly adhere to the standards of its industry, organization, or genre, such as title structure, order of sections, and use of punctuation? Attention to Detail : Are formatting details consistent throughout the document? Does the overall presentation reflect professionalism and standardization?

4. Logical Coherence and Persuasiveness

Clarity of Logic : Does the document exhibit a rigorous logical framework? Are the arguments connected by clear and explicit logical relationships? Persuasiveness : Does the document provide sufficient evidence or data to support its arguments? Does it effectively explain the background issues and propose reasonable solutions or viewpoints?

5. Adaptability and Goal Orientation

Contextual Relevance : Is the document tailored to specific contexts, target audiences, or time constraints? Does it align with the readers' expectations and needs? Clarity of Purpose : Does the document directly address its intended purpose? Is it clear and actionable enough to guide specific actions or communicate objectives effectively?

H.15 Speech

1. Clarity of Communication Goals

Core Message : Does the speech clearly establish its communication goal (e.g., to inform, persuade, or inspire)? Content Alignment : Does the content of the speech effectively support and achieve the intended goal? Conclusion and Guidance : Does the conclusion or call to action clearly guide the audience toward the desired action or thought?

2. Clarity and Logical Structure of Content

Key Points : Are the central ideas of the speech clear and easy to understand? Logical Organization : Is the speech logically structured, with smooth transitions between arguments? Conciseness : Does the content avoid ambiguity, unnecessary complexity, or overly obscure expressions?

3. Evidence and Support

Use of Facts and Data : Does the speech include relevant, reliable facts, data, or examples to support its claims? Sufficiency of Evidence : Is the provided evidence sufficient and convincing? Credibility of Information : Are the sources or evidence clearly cited to enhance the credibility of the information?

4. Depth and Relevance of Content

Depth of Analysis : Does the speech explore the topic in depth, avoiding overly superficial discussions? Audience Relevance : Does the content adequately consider the audience's interests, needs, and background, ensuring high relevance? Addressing Counterpoints : Does the speech anticipate potential concerns or opposing views from different segments of the audience, and respond appropriately?

5. Precision and Style of Language

Precision : Is the language used in the speech precise, avoiding ambiguity, wordiness, or unclear expressions? Style Appropriateness : Is the speech style suited to the topic and intended audience, with appropriate and respectful language? Clarity and Impact : Are the expressions concise and impactful, avoiding unnecessary information or repetition?

I Annotator Information

We hired 36 experts in writing with at least bachelor's degree and 23 of them are pursuing master degree or PhD degree in university. 29 of the experts major in literature, history, philosophy, journalism and communication, sociology, phychology and pedagogy. 7 of them are from engineering majors such as environment/engergy/computer science.

The pricing for each data is $10, containing 9 scoring assessment for 9 LLM writing.

J Completion Annotation Guidance

Completion Writing Scoring Criteria

I. Task Objectives, Fields & Techniques

A. Task Objectives

Assess the quality of responses filling the intermediate paragraph based on context, and score different responses. Responses A, B, and C are the model's completions for the text at the [fill in the blank] position. The reference completion is defined as a demonstration paragraph with a score of 4 points. You need to carefully read the context of the text needing completion and the reference completion, and score responses A, B, and C based on the specific dimensions provided in this rule.

B. Field Description

Fixed Fields (No annotation needed)

Instruction Content: Basic instruction requesting AI to fill in the blanks in the given text.

Text to be filled: The context with a missing intermediate part (emphasize careful reading), containing [fill in the blanks].

Reference Completion: The possible content to fill in the text, scored out of 5.

Responses A/B/C: The inferred missing context based on the instruction content and the partial text; these responses need to be scored later.

Note that replies may contain conversational content, which can be ignored, and only the fill-in content should be evaluated. If a response provides more than one fill-in example, only the first example should be evaluated. Annotated Fields (Fields you need to annotate) Each response has two annotation fields, where the scoring field is mandatory. Choose error types in the drop-down list for responses A/B/C as applicable.

Annotation Field 1: score A/B/C

Score the content format of response A/B/C based on the relevant rules in this document (e.g., instruction adherence, language expression, writing technique, emotional expression, writing style, etc.).

Annotation Field 2: Errors in Responses A/B/C (drop-down menu)

Note: This field is required if the score is below 3. Choose the relevant error type from the drop-down list (detailed error types can be found in the "2. Penalty Items - Error Types" section below).

C. Techniques / Points to Note

Thoroughly read the context around the [fill in the blank] to understand the writing logic.

It is recommended to use the computer screen split function to copy the text to be filled into http://annot.xhanz.cn/tools/markdown , then compare the reference completion and each model's response one by one.

Fact-check if there is factual content.

Accelerate the judgment process by referencing the "III. Scoring Basis (0) Scoring Logic" section.

II. Scoring Basis

Total score is 5 points, with the passing score being 3 points, and the minimum score being 1 point. The reference completion quality corresponds to a 4-point standard.

High-Quality Response: 4-5 points

Passing Response: 3 points

Low-Quality Response: 1-2 points

5 points: Quality surpasses the reference completion, meeting absolute dimension requirements (no penalty reasons).

4 points: Quality of content (language, logical emotional expression, etc.) is similar to the reference completion and meets absolute dimension requirements (no penalty reasons).

3 points: Meets absolute dimension requirements (no penalty reasons) but quality is lower than the reference completion (if there are penalty items, the score should be below 3).

2 points: 1-2 absolute dimensions are not met (requires penalty reasons).

1 point: (requires penalty reasons)

More than 2 absolute dimensions are not met;

Or, the response performs well in other dimensions (can be scored 3-5 points), but there is a severe security issue, or the [filling instruction] is not followed. In such cases, directly score 1 point.

Scoring Logic

Distinguish between high and low scores: First determine whether to score 1-2 points or 3-5 points based on the absolute criteria. For middle and high scores (3-5 points), assess based on the quality comparison with the reference completion.

For low scores (1-2 points), score 1-2 points based on penalty items and select the penalty reasons.

Finally, adjust to 1 point for responses with special issues (safety issues) and select the reason.

4-5 points Standard

4-5 points should be considered high-quality, comparable or better than the reference completion, from the following aspects:

Language Expression

Is the language more accurate and clear? Is the vocabulary more varied, making the description more vivid? Is the sentence structure more flexible, fitting the writing style better? Content Richness Does it appropriately cite speech, poetry, or allusions, adding cultural depth to the text? Writing Techniques/Artistic Presentation Are rhetorical devices used more aptly and skillfully?

Emotional Expression

Is the emotional expression more natural and forceful?

(A) Absolute Criteria (For a baseline score of 3) Up/Down Context Consistency: The completion should thoroughly comprehend and align with the context. Format: Consistent with preceding and following paragraphs. Content: Consistency in perspective/narrator Logical consistency Consistency in language style/tone Fact consistency: Any facts in the fill-in should logically align with the context if previously mentioned. Note: The fill-in isn't limited to an optimal reply (no need for the sole reference completion), only requiring coherent and logically consistent text. Accuracy: No factual errors in quoted external knowledge (publications, speeches, factual content). Fluency: The fill-in should be fluent, without language errors or logical contradictions, no mixed language issues, and no inappropriate use of special tags or numbering when not required.

(B) Penalty Items - Error Types

If the following errors are present, the score should be below 3.

A. Consistency Issues:

Format Inconsistency:

E.g., preceding or following paragraphs are long paragraphs while responses A/B/C are single sentences. Content Inconsistency: Inconsistent perspective/narrator Logical inconsistency Inconsistent language style/tone Repeated content: The fill-in should not reiterate context content. Score: 1-2 points deducted based on the severity. Notes: Different length from the reference isn't a penalty item.

B. Accuracy Issues:

Fact-check fill-ins for any factual errors. Need verification for: 1. Quoted statements 2. Published knowledge 3. Real-world place/company info 4. Concrete statistical data 5. Historical/news events 6. Facts for professional areas, like disease names. 7. Common sense mistakes, like the sun rising from the west. If factual errors are present, deduct 1-2 points based on the severity.

C. Fluency Issues:

1. Unmeaningful repetition.

Example: "Firstly... Secondly... Then..." shouldn't be used without necessity. Repeating or rephrasing the same point without deeper insight.

2. Mixed Language Issues. - Statements like "I say this is not okay" mixing languages deduct 2 points (score 1 point). - Clear English abbreviations that can be translated like "WC" to "toilet" deduct 1 point. - Common terms like "KFC" don't require translation, not a deduction item.

3. Special Character Issues. - Unfit characters, codes like "one, (1), ①" out of order or odd symbols like ˆ, &, deduct 1 point. Example of Errors: There are referencing and logic issues; if a part is repeated and an issue contextually misplaced, re-

sponses may score around 2 points as they fail to fit fill-in criteria aligned with reference points.

(C) Special Cases: Safety Issues (final step post scoring)

Directly score 1 point.

Generating violent, bloody, horrifying, obscene, or abusive content. Inducing self-harm, murder, societal revenge, or illegal content. Defamation against national leaders or governments. Incorrect representation of national leader's speeches.

K Evaluation Prompt Script Example

Example Evaluation Rubrics for Fictions

As a professional novel reviewer, please evaluate the following novel based on the provided criteria and scoring guidelines. For each dimension, assign a score from 1 to 10 and provide a brief explanation or justification for the score. Finally, give the novel an overall score on a scale from 1 to 10. A 6-point example will be provided beforehand for reference.

1. Plot and Structure Plot Compactness: ...... **Sense of Pacing**: ...... **2. Character Development** Depth of Characterization: .... Character Growth: .... Interpersonal Relationships: .... **3. Theme and Ideas** Thematic Depth: ... Expression of Ideas: ... Social or Cultural Context: ... **4.Language and Style** Language Style: ... Adaptability of Language: ... Detailing: ... **5.Emotional Resonance** Emotional Depth: ... Authenticity of Emotions: ... **6. Innovation and Uniqueness** Innovative Elements: ... Unique Perspective: ...

L LLM Prompts during Evaluation

L.1 Edge Weighting

Prompts for Edge Weighting

Input: Writing Instruction $\mathcal{I}$

Please assign a weight to each evaluation dimension based on the following writing instruction and evaluation dimensions. Follow these rules when assigning weights:
1. The sum of the weights of all evaluation dimensions must equal 1.
2. The weights should be floating-point numbers between 0 and 1, rounded to a maximum of two decimal places. In rare cases, negative weights are allowed but no lower than -1.
3. Each dimension's weight should be reasonably allocated according to its relevance to the characteristics of the writing instruction. Negative weights are permitted. For example, in argumentative writing, the weight for emotional expression can be set very low (e.g., 0 0.1) since emotional expression may hinder the rigor of argumentation.
4. After assigning weights to all dimensions, provide a brief explanation for your choices.
[Writing Instruction Start]
{instruction}
[Writing Instruction End]
[Evaluation Dimensions Start]
1. Introduction and Conclusion : The introduction should be engaging and innovative; the conclusion should go beyond mere summary, aiming to impress or resonate deeply, and avoid formulaic openings or endings.
2. Language and Rhetoric : Rich vocabulary and clear sentences; the writing should vividly describe objects (scenery, people, psychology, actions, etc.) and make skillful use of rhetoric or writing techniques (e.g., metaphor, parallelism).
3. Argumentative Logic : Logical progression should flow seamlessly, leading readers naturally from common knowledge to deeper thoughts; argumentation must be solid and avoid jumping to conclusions or excessive slogan-style assertions.
4. Emotional Expression : Tailored to the target audience and writing content, emotions conveyed by the author or characters should evoke strong resonance in readers.
[Evaluation Dimensions End]

M Implementation Prompts for ToW Experts

M.1 Opening and Ending

Opening and Ending Prompt

Input: Instruction, Reference, Content

Please act as a professional writing reviewer and evaluate the quality of the opening and closing sections of the "Writing to Be Evaluated" and the "Reference Writing." Your task is to analyze the strengths and weaknesses of the "Writing to Be Evaluated" based on the provided evaluation criteria and assign it a score between 1 to 10, along with a brief explanation of your reasoning.

The following content will be provided:

- **Evaluation Criteria**: Includes multiple dimensions and specific questions to help assess the quality of the opening and closing sections.

- **Writing Instructions**: The requirements, background, and main theme of the two pieces of writing.

- **Reference Writing and Writing to Be Evaluated**: Two pieces of writing excerpts to be compared.

Evaluation Criteria

A. Evaluation of Opening Quality

1. **Ability to attract the reader's attention**

- Does the opening grab the reader's attention and make them want to continue reading?

- Does it achieve this by using thought-provoking questions, engaging stories, shocking facts, or data?

2. **Clear introduction of the topic**

- Does the opening clearly convey the article's topic and direction?

- Does it establish the overall logical structure of the article, giving readers a clear expectation?

3. **Suitability for the target audience**

- Does the opening align with the target audience's interests or knowledge background?

- Is the language style suitable for the type of article (e.g., highly narrative for literary writing vs. precise for academic writing)?

4. **Avoidance of clichés or irrelevant content**

- Does the opening avoid overly common, flat, or dull phrasing?

- Does it go straight to the topic rather than being overly long or tangential?

5. **Appropriate emotional and atmospheric engagement**

- Does the writing create strong emotional impact or an engaging atmosphere (e.g., suspense, humor, tension)?
B. Evaluation of Closing Quality

1. **Summarization of core ideas**

- Does the conclusion clearly summarize the content of the article?

- Does it reinforce the central theme or idea, avoiding a "weak ending"?

2. **Deepening the theme**

- Does the conclusion help readers understand the significance or value of the article's message in greater depth?

- Does it elevate the argument through reflection, inspiration, or deeper insights?

3. **Leaving a strong impression or lingering effect**

- Does the conclusion evoke emotional resonance, thought, or motivate action?

- Does it end with a memorable sentence or concept?

4. **Structural and logical completeness**

- Does the conclusion echo the opening and the article's overall structure?

- Does it provide a natural sense of closure and avoid abrupt or rushed endings?

5. **Avoidance of excessive length or repetition**

- Is the conclusion concise and impactful, without repeating earlier details excessively?

- Does it avoid introducing new, unexplored points that disrupt the main thread of the article?

Writing Instructions

{instruction}
Reference Writing

{reference}
Writing to Be Evaluated
{content}
Evaluation Process

Please adhere strictly to the following steps to avoid contradictions:

1. **Strengths and Weaknesses Comparative Analysis**:

Using the evaluation criteria, analyze the performance of both the "Writing to Be Evaluated" and "Reference Writing" in terms of their opening and closing sections. Identify the relative strengths and weaknesses, ensuring detailed analysis across each criterion without omissions.

2. **Scoring and Reference Baseline**:

The “Reference Writing” is assigned a fixed baseline score of **6**, which serves as the standard for comparison. Based on the performance of the “Writing to Be Evaluated,” assign a score according to the following rules:

- **1-2 points**: The “Writing to Be Evaluated” is significantly weaker across nearly all evaluation criteria compared to the “Reference Writing.”

- **3-4 points**: The “Writing to Be Evaluated” is weaker in most evaluation criteria but slightly superior or equal in a few areas.

- **5-6 points**: The “Writing to Be Evaluated” shows a balanced performance compared to the "Reference Writing," being slightly better in certain aspects but generally equivalent overall.

- **7-8 points**: The “Writing to Be Evaluated” is stronger in most evaluation criteria compared to the “Reference Writing,” with only minor shortcomings.

- **9-10 points**: The “Writing to Be Evaluated” excels across nearly all criteria and demonstrates exceptional quality overall.

3. **Output Format**:

Please present the evaluation outcome in the following format:

Comparative Analysis:

1. Opening Section: Analysis content……

2. Closing Section: Analysis content……

Score: [[X]]

Reasons for the Score: ......

Important Notes:

- Summarize key points concisely while maintaining strict logical coherence.

- Use double square brackets (e.g., [[6]]) to indicate the score. Final scores must be whole numbers between 1 and 10.

Please proceed with the evaluation.

M.2 Metaphor

Metaphor Prompt

Input: Instruction, Reference, Content

Please evaluate the following "writing for evaluation" and "reference writing" in terms of language richness and appropriateness of rhetoric. Your task is to analyze the strengths and weaknesses of the writing for evaluation based on the provided assessment criteria, ultimately assigning a score from 1 to 10 and briefly explaining your reasoning.

You will be provided with the following items:

- Assessment Criteria: Guidelines on language richness and appropriate rhetoric, divided into positive and negative scenarios.
- Writing Instructions: Requirements, background, and main themes for both writings.
- Reference Writing and Writing for Evaluation: Two writing excerpts for comparison.

[Assessment Criteria Begins]
I. Positive Scenarios

1. Language Richness
- Use a diverse range of vocabulary, avoiding repetition or monotony, which demonstrates flexibility in written expression.
- Language exhibits expressiveness, precisely depicting and portraying scenes (e.g., landscapes, characters, psychological activities, actions), rendering the content vivid and lively.
- Crafted with meticulous attention to language, enhancing the cultural or artistic appeal of the writing through word choices.

2. Excellence and Appropriateness of Rhetoric
- Proper use of rhetorical devices contributes to vivid expression, depth of thought, or emotional impact. Examples include metaphors, repetition, parallelism, personification, etc.
- Rhetorical devices align with the logical flow of content, avoiding excessive embellishment and enhancing the power of communication.
- Writing techniques are not overdone, with rhetoric seamlessly integrated, blending naturally with the context and theme.

3. Structure and Logic
- The opening captures readers' attention with clear and compelling language and ideas.
- The conclusion is impactful and summarizing, elevating the main theme or inspiring further thought.

II. Negative Scenarios

1. Language Deficiency
- Monotonous vocabulary or overuse of repetitive words, such as frequent use of synonyms or simple words, making expression weak or immature.
- Sentences are poorly constructed, or grammatical errors noticeably affect the flow of reading.
- Generic or meaningless information dominates the content, e.g., vague discussions lacking specific details.

2. Inappropriate Use of Rhetoric
- Lack of rhetorical devices or reliance on only one type, resulting in overly flat or uninspired expression.
- Improper application of rhetorical devices, e.g., forced metaphors or overly complex

sentences, lowering the overall quality.
- Awkwardly inserted rhetoric disrupts content logic or diverges from the main theme.

3. Structural Problems Affecting Expression
- Overusing simple connectors such as "firstly, secondly, lastly," where the structure weakly relates to logical content.
- Failing to echo the main theme, e.g., conclusions that do not summarize critical points or openings that lack appeal.
- Expression limited to narrative progression or point listing, lacking deeper analysis or detailed depiction (e.g., bland storytelling or redundant argument repetition).

4. Mixed Language Issues
- Entirely mixed language styles, e.g., "我say 这个不行" or "highlight 了这页slide,” deduct 2 points.
- Unnecessarily retaining English abbreviations translatable into Chinese, e.g., using "WC" directly instead of translating, deduct 1 point.
- Conventional names like "KFC" and "NBA" not translated into Chinese are acceptable unless poorly matched with context.

5. Content Limitations in Argumentative or Fiction Writing
- [Fiction] Lack of detailed portrayal of characters or psychological depth, relying solely on narrative, merits deductions based on the importance of the plot.
- [Essay] Circular reasoning with no progressive analysis, e.g., listing pros and cons without further comparative summaries, merits 1-2 points deduction.

[Assessment Criteria Ends]

[Writing Instructions Begin]
{instruction}
Writing Instructions Ends

[Reference Writing]
{reference}
Reference Writing Ends

[Writing for Evaluation Begins]
{content}
Writing for Evaluation Ends

Follow the steps below to complete the evaluation. Avoid contradictory logic:

1. Strengths and Weaknesses Comparative Analysis: Compare the performance of the "writing for evaluation" and "reference writing" based on the assessment criteria, analyzing relative strengths and weaknesses step by step.

2. Scoring Based on Reference Benchmark: The fixed score for the reference writing is 6 points, which serves as the baseline. Use the scoring guidelines below to judge the writing for evaluation:
1-2 Points: Writing for Evaluation exhibits significantly more negative scenarios in comparison to Reference Writing, with nearly no positive scenarios, appearing overly simplistic or lacking examples.
3-4 Points: Writing for Evaluation exhibits slightly more negative scenarios in comparison to Reference Writing, showing fewer positive scenarios, with content lacking vividness or convincing argumentation.
5-6 Points: Writing for Evaluation exhibits positive and negative scenarios on par with Reference Writing, demonstrating average performance, with some use of rhetoric, techniques, or examples, albeit less naturally.
7-8 Points: Writing for Evaluation exhibits similar negative scenarios to Reference Writing but features more positive scenarios, with language flowing naturally and enriched with expression or examples.
9-10 Points: Writing for Evaluation exhibits significantly more positive scenarios compared to Reference Writing, with little to no negative scenarios, demonstrating rich, vibrant, and well-balanced expression while maintaining logical progression.

3. Output Format: Deliver the evaluation in the following format:
"‘ Strengths and Weaknesses Analysis:
1. Positive Scenarios Comparison: Analysis content. . . . . .
2. Negative Scenarios Comparison: Analysis content. . . . . .

Score: [[X]]

Score Reasoning:. . . . . .
"‘

Notes:
- Summarize key points succinctly, ensuring logical consistency.
- The "Score" section must encapsulate the number in double square brackets (e.g., [[6]]). The final score must range between 1 and 10 inclusively.

Begin your evaluation.

M.3 Logics

Logics Prompt

Input: Instruction, Reference, Content

Translate the instructions into English below, using double slashes ‘
‘ to separate all new lines:
—
Please act as a professional writing reviewer and evaluate the reasoning logic in the "Writing to Be Evaluated" and "Reference Writing" provided below. Your task is to analyze the strengths and weaknesses of the "Writing to Be Evaluated" based on the evaluation criteria, provide a score between 1 and 10, and briefly explain your reasoning.

You will receive the following:
- Evaluation Criteria: Descriptions regarding logical reasoning, categorized into ideal scenarios and poor performances.
- Writing Instructions: Requirements, context, and main themes of the two pieces of writing.
- Reference Writing and Writing to Be Evaluated: Two writing fragments for comparison.

[Evaluation Criteria Starts]
The core focus of content logic:
1. Consistency in person/point of view: Does the entire piece maintain a consistent narrative style and perspective, avoiding abrupt shifts? If there are changes in person or perspective, is there prior groundwork or subsequent explanation?
2. Logical coherence: Is the reasoning process internally consistent? Do the ideas naturally connect without abrupt breaks or gaps in logic?
3. Consistency in language style and tone: Does the expression retain consistent tone and style throughout, ensuring a smooth and natural reasoning process?
Note: Reasoning logic typically pertains to issues within a paragraph or across a few adjacent paragraphs. The focus is on maintaining context continuity and logical consistency.

1. Ideal Scenarios (Exemplars of Excellent Reasoning Logic):
1. Logical reasoning in argumentative writing:
- The reasoning process is tightly interconnected, with clearly defined logical layers. Ideas progress naturally from common knowledge to deeper analysis, enabling readers to follow step-by-step reasoning.
- The content includes both abstract theoretical analysis and concrete evidence to support conclusions, forming naturally persuasive arguments.

2. Structure in speeches or addresses:
- The logic is clear and straightforward: identify the core issue, present viewpoints for addressing the issue, then explain steps or solutions with specificity.
- Ideas transition from macro-level problem analysis to specific actionable methods, culminating in an inspiring conclusion with layered content.

3. Writing in application letters or summary

reports:
- Reports unfold content logically following a structure like "Objective → Problem Analysis → Key Challenges → Solutions → Achieved Results," reflecting clear work thought processes and reasoning.
- Application letters discuss the attributes and significance of the requested entity while precisely aligning it with the applicant's needs, forming a tight connection and enhancing the reasonableness and persuasiveness of the content.

2. Poor Performances (Issues and Deduction Standards):
If a piece displays the following logical problems, scores should be deducted accordingly:

1. Issues in argumentative writing:
- Content solely lists opinions without reasoning or evidence, such as "We should XXXX, we must XXXX," lacking justification, reasoning, or examples. This glaring lack of logic warrants at least a 3-point deduction.

2. Problems in speeches or addresses:
- Analysis or solutions lack depth or broadness. For example, overly abstract discussions without actionable plans, or overly specific ideas without high-level thought enhancements. Deduct 3-6 points based on severity.

3. Flaws in application letters or summary reports:
- Work summaries only state actions performed without analysis or reasoning, making the content superficial.
- Application letters fail to establish alignment between the requester's needs and the requested entity, resulting in vague or disconnected content. Deduct 3-6 points based on the severity of issues.

[Evaluation Criteria Ends]

[Writing Instructions Starts]
{instruction}
Writing Instructions Ends

[Reference Writing]

{reference}

[Reference Writing Ends]

[Writing to Be Evaluated Starts]
{content}
Writing to Be Evaluated Ends

Please strictly follow the steps below to complete the evaluation, avoiding contradictions in logic:
1. Comparative Analysis of Strengths and Weaknesses: Compare the performance of "Writing to Be Evaluated" against "Reference Writing" based on the evaluation criteria, gradually analyzing their relative strengths and weaknesses across all points.

2. Scoring with Reference Baseline: Use the "Reference Writing" as the 8-point standard of excellence. Start with 8 points for the "Writing to Be Evaluated" and deduct scores as follows:

Deduction Rules:
- For 0-1 minor weakness or logical flaw, deduct 0-3 points.
- For 1-2 minor weaknesses, deduct 3-6 points.
- For 2 or more minor weaknesses, or 1 significant logical problem, deduct 6-8 points. Severe issues may justify further deductions.

Additional Points:
- If the evaluated writing demonstrates clear advantages over the reference writing, you can award additional points, but no more than 2 points.

Other Notes:
- A score between 1-4 does not necessarily mean poor reasoning logic. Simply base your score on the weaknesses analyzed.

3. Output Format: Use the following format

for your evaluation result:

‘‘‘
Strengths and Weaknesses Analysis:
1. Comparison of positive aspects: Analysis content. . .
2. Comparison of negative aspects: Analysis content. . .

Score: [[X]]

Reason for score: . . .
‘‘‘

Notes:
- Briefly summarize key points while ensuring logical rigor.
- The “Score” section must use double brackets to enclose the number (e.g., [[6]]). The final score must be a whole number between 1 and 10.

Please begin your evaluation.

M.4 Emotion

Emotion Prompt

Input: Instruction, Reference, Content

Please act as a professional writing reviewer and evaluate the "Writing to Evaluate" and "Reference Writing" for emotional expression quality. Your task is to analyze the strengths and weaknesses of the "Writing to Evaluate" according to the given evaluation criteria. Then, provide a score between 1 and 10 and briefly explain your reasoning.

You will be given the following content:

- Evaluation Criteria: Guidelines on emotional expression, divided into good and poor examples.

- Writing Instructions: The requirements, background, and main themes of the two pieces of writing.

- Reference Writing and Writing to Evaluate: Two writing excerpts for comparison.
Emotional expression must connect to the intended readers, meaning it should consider target readers before conveying emotions.

Good Conditions:

1. Emotion is successfully integrated into various descriptions (e.g., events, scenery, character portrayals), making the writing warm and layered, while enhancing emotional tension.

2. Skillful use of rhetorical devices such as metaphor or personification improves emotional expressiveness, creating greater visual appeal and emotional impact.

3. Tone, vocabulary, and sentence structures match the target reader’s style. The emotions are genuine and fluid, capable of resonating deeply with readers or sparking contemplation.

4. Emotional expression aligns closely with the main theme, helping to drive the narrative forward or deepen core points.

Poor Conditions: 1. Ineffective or missing emotional expression that fails to convey the intended feeling or is disconnected from the emotional context.

2. Misuse of rhetorical devices, e.g., improper metaphors or unrelated comparisons that weaken emotional expression.

3. Emotional expression is superficial, exaggerated, or unnatural, making it difficult for readers to truly relate or empathize.

{instruction}
Writing Instructions End

{reference}
Reference Writing End

{content}
Writing to Evaluate End

Please strictly follow the steps below and avoid contradictory logic:

1. **Strengths and Weaknesses Analysis:** Analyze the relative strengths and weaknesses of the "Writing to Evaluate" compared to the "Reference Writing" based on the evaluation criteria, systematically addressing each point.

2. **Scoring and Reference Baseline:** The Reference Writing is fixed at a score of 6, which serves as the baseline for comparison. Assign a score to the Writing to Evaluate using the following scoring rules:

- **1-2 Points:** Writing to Evaluate displays significantly more poor conditions than the Reference Writing. It lacks good conditions, leading to monotonous, shallow text with weak emotional portrayal or insufficient author sentiment.

- **3-4 Points:** Writing to Evaluate displays somewhat more poor conditions than the Reference Writing, and fewer good conditions. The text has some emotional portrayal but is mediocre, or contains abrupt or unsuitable elements.

- **5-6 Points:** Writing to Evaluate showcases a balance of good and poor conditions similar to the Reference Writing. It performs adequately and has some level of reader emotional engagement.

- **7-8 Points:** Writing to Evaluate demonstrates an emotional quality close to the Reference Writing, but with additional good conditions. It portrays a variety of emotions that complement each other or utilizes techniques like environmental descriptions to convey emotions appropriately.

- **9-10 Points:** Writing to Evaluate exhibits significantly more good conditions than poor ones, displaying rich and nuanced emotions. Readers are deeply moved by the author's sentiment, or emotions are skillfully conveyed through elements like environmental descriptions.

3. **Output Format:** Use the following format for the evaluation result:

Strengths and Weaknesses Analysis:

1. Comparison of Good Conditions: Analysis content...

2. Comparison of Poor Conditions: Analysis content...

Score: [[X]]

Reason for the Score: ...

Note:

- Concisely summarize key points, ensuring logical rigor.

- Use double brackets ([[X]]) to denote the final score, which must be an integer between 1 and 10.

You may now proceed with the evaluation.

M.5 Plots

Plots Prompt

Input: Instruction, Reference, Content

Please act as a professional writing evaluator and assess the appropriateness of the plot design and development in the following “Writing to be Evaluated” and “Reference Writing.” Your task is to analyze the strengths and weaknesses of the “Writing to be Evaluated” based on the provided evaluation criteria, then assign a score from 1 to 10, briefly explaining your reasons.

You will receive the following content:

- Evaluation Criteria: Explanation of plot design and development, divided into good and bad scenarios.

- Writing Instructions: Requirements, background, and central idea for both pieces of writing.

- Reference Writing and Writing to be Evaluated: Two writing samples for comparison.

[Evaluation Criteria Start]

I. Good Scenarios

1. Structure Logic:

- Overall structure is clear, well-organized, and logically sound.

- Paragraph transitions are smooth and coherent, with lively inter-sentence transitions instead of relying on mechanical connectors (such as “firstly, secondly, then”).

- The plot unfolds in an orderly manner, with internal consistency, foreshadowing that is later resolved or explained.

2. Narrative Techniques:

- Skillful use of techniques such as flashbacks or insertions, resulting in a concise narrative with clear focus.

- Well-paced progression, ability to create tension at key plot points, capturing reader interest or building suspense.

- Narrative logic is reasonable, with causality or thematic guidance; plot does not appear abrupt or forced.

3. Thematic Value:

- Thematic content is novel and meaningful, able to resonate with readers or provoke deeper thought.

- The narrative conveys insightful perspectives or elevated value orientations.

4. Fit to Writing Type:

- Argumentative essay: Clear thesis, well-organized points, sufficient and relevant evidence.

- Fiction: Characters and plot serve the theme; narrative is tight and vivid.

- Prose: Beautiful imagery, unified theme, lively and evocative language.

- Speech, Application, Report: Concise, focused, logically flowing, persuasive.

II. Bad Scenarios

(Evaluation basis: Minor issues deduct 3 points; significant or severe problems deduct 6 points)

1. Structure & Organization Issues:

- The structure is disorganized with no hierarchical progression.

- Paragraphs or sections lack transitions and connections, with abrupt jumps or content piling up, leading to confusion.

- Plot progression is flat and tedious, lacking attraction or logical coherence.

2. Narrative Issues:

- Improper use of flashbacks/insertions, resulting in confusion or breaking reading immersion.

- Plot design is loose or fragmented, failing to establish tension or draw interest.

- Plot development lacks causality or necessary foreshadowing, with unconvincing or abrupt logic.

3. Thematic Issues:

- The theme is cliché, dull, or lacks originality, failing to engage or resonate with readers.

- The intention remains superficial, lacking depth or meaningful development.

- Lacks thematic unity; content is scattered, making it hard to consolidate into an overall viewpoint or main idea.

4. Appropriateness Issues:

- Argumentative essay: Thesis deviates from topic, points are simplistic or drawn out, evidence is insufficient or weak.

- Fiction: Characters are stiff, lack appeal; plot design has no bearing on theme or is disconnected from character behavior.

- Prose: Imagery is bland, lacks vividness; fragmented content fails to support the central theme.

- Speech, Application, Report: Sentences are verbose or too casual, lack focus, logical disconnections.

5. Language Issues:

- Language is simplistic or monotonous, lacking vitality and expressiveness.

- Writing is obscure or hard to understand, or the style does not match the intent.

- Excessive use of mechanical or formulaic sentences, lacking writer's individual voice.

[Evaluation Criteria End]

[Writing Instructions Start]
{instruction}
Writing Instructions End

[Reference Writing]
{reference}
Reference Writing

[Writing to be Evaluated Start]
{content} [Writing to be Evaluated End]

Please strictly follow these steps for evaluation and avoid contradictory logic:

1. Comparative Analysis: Using the evaluation criteria, analyze the strengths and weaknesses of the "Writing to be Evaluated" in comparison to the "Reference Writing," covering all relevant points.

2. Scoring and Reference Baseline: The Reference Writing is assigned a fixed score of 6, serving as the baseline for assessment. Based on the performance of the Writing to be Evaluated, use the following rules to determine your score:

1-2 points: The Writing to be Evaluated displays clearly more bad than good elements compared to the Reference Writing; the narrative is largely a dry recounting, lacking linguistic richness and rhetorical depth, with thin content and weak logic, and significant flaws.

3-4 points: The Writing to be Evaluated displays somewhat more bad than good elements than the Reference Writing; narrative is plain and rather unengaging, plot is relatively bland, with weak rhetorical and detailed description, but the overall logic is basically coherent.

5-6 points: The Writing to be Evaluated and Reference Writing have roughly equivalent numbers of good and bad elements; language is mostly fluent, plot is complete with no major flaws, structure is unified with some internal cohesion, but highlights are lacking.

7-8 points: The Writing to be Evaluated displays a similar number of bad elements as the Reference Writing, but clear strengths as well. There are highlights in language or plot, such as successfully catching readers' attention, engaging twists, tightly structured and logical progression, and a layered and impactful narrative.

9-10 points: The Writing to be Evaluated displays obviously more good elements than the Reference Writing. The plot is full of tension, logic is rigorous and concise, with distinctive strengths in multiple aspects, demonstrating high-level writing and artistry.

3. Output Format: Please present your evaluation in the following format:

“‘ Strengths and Weaknesses Analysis:
1. Comparison of Good Elements: Analysis...
2. Comparison of Bad Elements: Analysis...

Score: [[X]]

Reason for Score: ...
“‘
Notes:

- Briefly summarize key points, ensuring logical clarity.

- The “Score” should be wrapped in double square brackets, e.g., [[6]]. The final score must be an integer between 1 and 10.

Begin your evaluation.

M.6 Paragraphing

Paragraphing Prompt

Input: Instruction, Reference, Content

Please evaluate the rationality of chapter and paragraph division in the writing below according to the following principles:

1. Coordination between sections: Try to avoid overly long paragraphs exceeding 500 characters or a chapter being disproportionately longer than all other chapters combined. Ideally, the length of each section should not vary significantly.

2. Necessity of chapter division: Dividing content by chapters using titles is necessary for texts with strong structural requirements, such as lengthy novels, reports, or speeches. Avoid meaningless chapter division, such as splitting text into separate chapters when it could have been merged into one. Ideally, each chapter should aim for a word count exceeding 800 characters.

3. Necessity of paragraph division: All writing aside from poetry should avoid excessive paragraphing, such as having a majority of paragraphs consisting of only one sentence. For poetry, paragraphs can be divided according to rhythm. Additionally, avoid not segmenting thoughts, where too many themes or arguments are lumped into a single paragraph. Ideally, each paragraph should exceed 50 characters, though brief paragraphs can be used for reasons like rhythm or tone.

Based on the three points above, score the text as follows:

- 3 points: Division is reasonable.

- 2 points: Division is somewhat flawed in chapters or paragraphs.

- 1 point: Division is clearly uncoordinated, lacks structured planning, or evaluation content is missing.

[Writing Instructions Start]
{instruction}
Writing Instructions End

[Writing for Evaluation Start]
{content}
Writing for Evaluation End

After assigning a score, provide a brief explanation of the rating.
Provide your review enclosed in double square brackets, such as "Overall score: [[2]]. Reason...".

The final scores are projected to a 1-10 scale by

$y = 5 \times (x - 1)$

M.7 Impression

Impression Prompt

Input: Instruction, Reference, Content

Please provide an overall impression score based on the following writing instructions and content. You may focus on the attractiveness of the beginning and ending, the fluency and coherence of the text, the appropriateness of the paragraph structure, and the avoidance of excessive use of bullet points and lists.

[Writing Instructions Start]
{instruction}
Writing Instructions End

[Reference Writing Start]
{reference}
Reference Writing End

[Writing for Evaluation Start]
{content}
Writing for Evaluation End

Please provide an integer overall impression score between 1 and 10, and briefly explain your reasoning. You need to format the score within double square brackets, e.g., "Overall score: [[8]]. Reason...".

M.8 Metaphor

Metaphor Prompt

Input: Instruction, Reference, Content

Please assign weights to each evaluation dimension based on the following writing instructions and evaluation dimensions. Below are the rules for assigning weights:

1. The total weight of all evaluation dimensions must equal 1.

2. Each weight must be a floating-point number between 0 and 1, rounded to no more than two decimal places.

3. Each dimension's weight should be reasonably distributed based on the relevance of the dimension to the characteristics of the writing instructions. Negative values are permitted. For example, in highly expository writing, the emotional expression weight can be set very low (e.g., 0 0.1) since emotional expression can hinder the logical rigor of the discussion.

4. After assigning weights to all dimensions, provide a brief explanation for your decisions.

[Writing Instructions Start]
{instruction}
Writing Instructions End

[Evaluation Dimensions Start]

Opening and Closing: The opening should capture attention and be original; the closing should not merely summarize but leave a lasting impression, reveal the article's deeper meaning, or evoke reflection. Avoid formulaic openings and endings.

Language and Rhetoric: Use rich vocabulary and clear sentences. Skillfully employ descriptive language (for scenery, characters, psychology, actions, etc.) to vividly depict subjects. Make use of brilliant rhetorical devices and writing techniques (e.g., metaphor, parallelism, etc.).

Reasoning Logic: Establish a flow of logic that seamlessly connects ideas, making the writing feel natural. Start with common knowledge and lead readers into deeper reflections step by step. Ensure arguments are solid and avoid abrupt jumps to conclusions or slogan-like assertions.

Emotional Expression: Based on the target

audience and writing content, convey the author's emotions or depict characters' feelings to evoke strong empathy in readers.
Evaluation Dimensions End

N Data Examples for Each Tasks

We show three examples for Completion, Guide, Open tasks in Table 12, 14, 16. Their English translations are presented in Table 13, 15, 17.

Setting	Prompt
Instruction	请根据上下文补全以下文章中用 [fill in the blank] 特殊符号标记出的内容。
Information	最近，有媒体盘点出了中国超级工程里的“世界之最”，在网络上引发了一大波热议： 白鹤滩水电站是目前世界在建规模最大、技术难度最高的水电工程；港珠澳大桥是世界上总体跨度最长的跨海大桥；新疆和若铁路开通运营，让世界首条环沙漠铁路线完成“最后一块拼图”…… 有网友称，“中国制造就是中国骄傲”。 而如果我们往深了扒一扒，超级工程的背后，实际上凝聚了大量自主研发的新科技，科技自强自立的背后，最终则是创新的驱动。 习近平总书记在党的二十大报告中22次提到创新，并深刻指出：坚持创新在我国现代化建设全局中的核心地位。报告中还有一处提到：创新是第一动力。 我们来理一理，“核心地位+第一动力”，创新的分量为何这么重? 一 从人类历史来看，社会生产力的每一次发展、科学技术的每一次进步，无不是通过创新实现的。 欧美几个发达国家就是抓住了科技和产业革命的创新机会而一跃跨入现代化行列，实现大国崛起和民族振兴，并引领时代的走向和世界的发展。 有创新就会有发展，谋创新就能谋未来。涅于一穷二白旧社会的中国式现代化，也经历了无数次以创新求发展的浴火重生。 特别是新时代以来，在创新驱动发展战略的指引下，我国的“创新型国家”的建设稳步加快。从2012年到2021年，全社会研发投入从1.03万亿元增长到2.79万亿元，全球创新指数排名从第34位上升到第12位。 科技创新在企业壮大、产业升级、区域发展、重大工程建设等方面发挥了重要作用，有力支撑了高质量发展，带动一些关键核心技术相继实现突破，取得了载人航天、探月探火、深海深地探测、超级计算机等重大成果。 九天之外传来的“感觉良好”，深潜海底万米的“妙不可言”，乘坐“复兴号”飞驰万里，睁开“天眼”仰观浩渺宇宙......这些，都成了网民心中中国科技创新的“名场面”，成了我们心中升腾起的自信和自豪。 二 [fill in the blank] 这样的故事还有不少。这些年，我们在科技“从模仿到创新”的转型过程中遭遇了“追赶的极限”，关键领域核心技术被“卡脖子”的问题愈发突出。 特别是中美贸易摩擦中，我国“缺芯少核”的科技短板暴露了出来。美西方国家利用技术优势地位一方面禁止关键技术流入中国，推动高科技产业链的“对华脱钩”；另一方面阻碍我国核心技术研发，企图将我国彻底压制在产业链中低端。 在激烈的国际竞争中，惟创新者进，惟创新者强，惟创新者胜。正是因为我国科技实力和世界领先水平的差距在不断缩小，一些领域实现了从“跟跑”到“并跑”甚至“领跑”，才引发了美西方国家的战略焦虑，并招致不惜成本的封锁和打压。 然而，我们的目标绝不是跟着西方国家亦步亦趋。我们要开拓出中国式现代化路径，这是一条从未有人走过的路。为人类实现现代化提供新选择，科技创新在其中的核心作用无疑更加凸显。 三 东部沿海省份浙江，为创新之路探了路。 早在2006年，习近平同志在浙江工作时就为浙江定下了用15年时间进入创新型省份行列，基本建成科技强省的目标。当年的“全省科学技术大会”这个会议名称，被习近平同志修改为“全省自主创新大会”。几字之变，意图更加清晰，导向更加明确。 一路走来，“自主创新”这面旗帜始终在之江大地上高高飘扬。今天的浙江，已经拥有良好的科创环境和氛围，三大科创高地加速打造。很多人一提到科创大走廊、之江实验室、西湖大学就想到浙江，这些高能级的平台不仅是浙江的“标签”，也正成为创新的沃土。 有活力就有人才，浙江也越来越成为顶尖人才的向往之地。截至今年8月，全省研发人员总量已达77.58万人，这就意味着大概每1000个浙江人中就有12个科研人员。 而这些科研平台、科研技术、创新力量，则前所未有地融入到百姓的日常当中。在全国率先启动数字化改革一年多来，浙江打造出一批实用、管用的重大应用。“海外智慧物流”“浙农服”“健康码”“政采云”……一个个有着鲜明浙江烙印的数字化应用，便企惠民，香飘墙外、飞向万家。 每个时代，都有打开创新之门的钥匙。比如第一次工业革命是蒸汽机，第二次工业革命是电气化。今天，浙江则以“数”谋“新”，做第一个吃螃蟹的人。 四 今天的世界瞬息万变。大变局之下，唯一的“不变之道”就是以变应变、以新应变。 创新，该怎么创？如何新? “必须坚持科技是第一生产力、人才是第一资源、创新是第一动力，深入实施科教兴国战略、人才强国战略、创新驱动发展战略，开辟发展新领域新赛道，不断塑造发展新动能新优势。”党的二十大报告中的这段话，为创新之路擘画了清晰的领域和路径。 此外，笔者认为，以创新驱动发展还要坚持好以下几个关键点。 创新靠不得别人，还得靠自己。创新能力，关乎一个国家在世界格局中的地位，甚至关乎着国家安全。在世界竞技赛中，跟着别人跑随时可能会被绊倒，只有把创新的自主权、技术的所有权、发展的主动权紧紧攥在自己手中，才能跑出速度、跑到前列。 创新的重要目的之一，是整合资源，打通链条、畅通循环。中国已经是全球第二大经济体，依靠传统的土地、资源和低成本人力来驱动发展已经没有竞争力，也不会有出路。只有用好新型举国体制优势，发挥创新的核心作用，打通不受制于人的产业链、供应链，才能在经济发展中涌现出无数“风口”，在国际竞争中站稳脚跟。 真正的创新，最终要落脚于民。近年来，我国科技创新能力不断提升，越来越多的创新成果广泛应用于民生领域。高铁网络、电子商务、移动支付、互联网+、共享经济……正在深刻改变着人们的衣食住行。不过，实现“人的现代化”也还有很多空白领域，如何围绕老百姓的切身需要，填补这些空白，是需要瞄准的“靶子”。 赶考路上，需要创新来“澎湃”。坚持创新在我国现代化建设全局中的核心地位，坚持创新是第一动力，不仅要让1不断地递增出N，也要探索如何让更多的0实现1的突破。
Reference	科技创新是大国竞争的核心领域。一个国家科技创新能力的高低，决定了其在国际竞争中的水平。 一个经典的故事是，1960年前后，一套重量为3公斤的精密光学坐标镗床主轴轴承，外商对我们的要价竟相当于和轴承同等重量的黄金或6吨重的对虾。直到我们通过自主创新成功攻关，才不再需要依赖进口。 这至少告诉我们两个道理： 第一，关键核心技术要不来、买不来、讨不来。只有把它牢牢攥在自己手中，才能从根本上保障国家总体安全。 第二，在现代世界体系中，不同国家有着不同的分工。位于“中心地区”的发达国家享有先进技术和高附加值产业，而位于“边缘地区”的欠发达国家只能提供原材料、自然资源和廉价劳动力。这一格局让资本和价值源源不断地向“中心地区”聚集并导致严重的两极分化。

Table 12: Example for **Completion** in Chinese.

Setting	Prompt
Instruction	Complete the contents whose position is marked with [fill in the blank] according to contexts.
Information	Recently, the media compiled a list of “world’s best” super projects in China, sparking lively discussions online: Baihetan Hydropower Station is currently the largest under-construction hydropower project in the world, with the highest technical difficulty; the Hong Kong–Zhuhai–Macau Bridge is the longest cross-sea bridge in the world; the opening and operation of the Xinjiang Hotan-Ruoqiang Railway has completed the “last piece of the puzzle” for the world’s first desert-circling railway line... Some netizens remarked, "Made in China is China’s pride." However, when we dig deeper, we find that behind these super projects lie significant new technologies developed independently, backed by the drive for technological self-reliance and self-strengthening, which in turn is fueled by innovation. In the report to the 20th National Congress of the Communist Party of China, General Secretary Xi Jinping mentioned innovation 22 times and profoundly emphasized: Innovation must occupy the core position in China’s overall modernization strategy. The report also stated: Innovation is the primary driving force. Let’s unpack this—“core position + primary driving force.” Why does innovation weigh so heavily? **I** In human history, every advancement in social productivity and every progress in science and technology has always been achieved through innovation. Several developed Western countries, such as those in Europe and North America, managed to seize the opportunities brought by technological and industrial revolutions, propelling themselves into the ranks of modernized nations, achieving national rejuvenation and rise to prominence, and leading the trajectory of their times and global progress. Where there is innovation, there is development; where there is a plan for innovation, there is a plan for the future. China’s modernization, which rose from a once-impoverished and backward society, has also undergone countless “phoenix-like rebirths” driven by innovation to seek development. Especially since the advent of the new era, under the guidance of the innovation-driven development strategy, China has been steadily accelerating its progress as an “innovative nation.” From 2012 to 2021, nationwide R&D expenditures increased from 1.03 trillion yuan to 2.79 trillion yuan, and the global innovation index ranking rose from 34th to 12th. Technological innovation has played a vital role in driving business growth, industrial upgrades, regional development, and the construction of major projects. It strongly supports high-quality development, enabling breakthroughs across critical core technologies in areas such as manned spaceflight, lunar and Mars exploration, deep-sea and deep-earth exploration, and supercomputers. The “feeling good” phrase transmitted from outer space, the “beyond words” achievement of deep-sea dives exceeding 10,000 meters, the miles sped through on the “Fuxing” bullet train, and the vast universe explored using the “Sky Eye”... All these iconic moments of China’s technological innovation have captured netizens’ imaginations, igniting pride and confidence in all our hearts. **II** [fill in the blank] There are many more stories like this. In recent years, during China’s transformative journey from “imitation” to “innovation,” we have encountered the “limits of catching up,” with challenges in critical core technologies increasingly coming to the forefront. Particularly during the U.S.-China trade friction, the technological shortcomings labeled as China’s “chip deficiency” and “lack of core technologies” were laid bare. Western countries, leveraging their technical dominance, simultaneously imposed bans on transferring critical technologies to China and tried to “decouple” high-tech industrial chains from China. They also sought to obstruct China’s R&D of core technologies in an attempt to suppress China to the lower ends of the industrial chain. In the fierce international competition, only those who innovate advance, only those who innovate become stronger, and only those who innovate win. It is precisely because the gap between China’s technological strength and world-leading levels is narrowing, with some fields accomplishing shifts from “running behind” to “running alongside” or even “leading,” that strategic anxiety has arisen among Western countries, prompting them to resort to cost-no-object blockades and suppression. However, our goal is not to follow in the footsteps of Western nations. Our aim is to pioneer a uniquely Chinese path to modernization—a road never before taken. Offering humanity an alternative modernization model makes the core role of technological innovation even more prominent. **III** The eastern coastal province of Zhejiang has been a trailblazer in the journey of innovation. Back in 2006, while working in Zhejiang, Comrade Xi Jinping set the goal of making Zhejiang an innovation-oriented province within 15 years and essentially building it into a province strong in science and technology. The conference, originally named the “Provincial Science and Technology Conference,” was renamed by Xi Jinping as the “Provincial Independent Innovation Conference.” This subtle change in wording carried a clearer intent and a more focused objective. Over time, the flag of “independent innovation” has flown high across the land of Zhejiang. Today, Zhejiang boasts an excellent environment and atmosphere for scientific and technological innovation, with three major innovation centers being rapidly developed. Mentioning the Innovation Corridor, the Zhijiang Laboratory, or Westlake University immediately brings Zhejiang to mind. These high-caliber platforms are not only among Zhejiang’s prominent “labels” but are also becoming fertile ground for innovation. Where there is vitality, there is talent. Zhejiang has increasingly become a magnet for top-tier talent. As of this August, the total number of R&D personnel in the province had reached 775,800, meaning that approximately 12 out of every 1,000 people in Zhejiang work in research. These research platforms, technologies, and innovation resources have also been unprecedentedly integrated into the daily lives of ordinary people. Thanks to Zhejiang’s bold steps in launching its digitization reform efforts, practical and user-friendly applications have emerged, such as “Overseas Smart Logistics,” “Zhejiang Agricultural Services,” “Health Code,” and “Government Procurement Cloud.” These digital services, bearing the unmistakable imprint of Zhejiang, have benefited businesses and citizens alike, extending their influence far beyond the region. Every era has its own key that unlocks the door to innovation. For instance, the steam engine in the First Industrial Revolution and electrification in the Second Industrial Revolution. Today, Zhejiang is creating the new with “data,” becoming the first to “try new things.” **IV** The world today is undergoing rapid changes. In an age of great transformations, the only “constant” is to adapt to change with change, and to respond to the new with the new. How should we innovate? What constitutes “new”? “We must uphold the principle that science and technology are the primary productive forces, talent is the primary resource, and innovation is the primary driving force. We must intensively implement the strategies of rejuvenating the country through science and education, strengthening the nation through talent, and driving development through innovation. We must continuously open new fields and tracks for development and create new momentum and new advantages for growth.” This excerpt from the 20th National Congress report outlines a clear roadmap for the path of innovation. Moreover, the author believes that driving development through innovation requires adhering to the following key principles: Innovation cannot rely on others; it must depend on ourselves. Innovation capacity determines a nation’s standing in the global landscape and even its national security. In the global competition arena, following others always carries the risk of being tripped. Only by firmly grasping the autonomy of innovation, ownership of core technologies, and initiative in development can we achieve speed and move to the forefront. One of the primary objectives of innovation is to integrate resources, streamline the chain, and ensure smooth circulation. As the world’s second-largest economy, China can no longer depend on traditional drivers such as land, resources, and low-cost labor for competitiveness or growth. By utilizing the advantages of the new nationwide system and emphasizing the role of innovation, China can build an autonomous and robust industrial and supply chain, generate numerous “opportunities” for economic growth, and secure its position in international competition. True innovation must ultimately focus on people. In recent years, China’s technological innovation prowess has steadily improved, leading to the widespread application of many innovative achievements in the realm of public welfare. High-speed rail networks, e-commerce, mobile payments, Internet+, the shared economy... all these have profoundly transformed people’s livelihoods. However, there remain many gaps in achieving “human modernization.” Addressing these gaps and meeting the genuine needs of ordinary people becomes the target to aim for. On this challenging journey, innovation serves as the driving force that powers us forward. By maintaining innovation as the core position in China’s modernization strategy and upholding it as the primary driving force, we must not only ensure the continuous transformation of 1 into N but also explore how to turn more zeros into breakthroughs of 1.
Reference	Technological innovation is the core arena of competition among major powers. The level of a country’s capacity for technological innovation determines its standing in international competition. A classic story goes that, around 1960, a set of precision optical coordinate boring machine spindle bearings weighing 3 kilograms was offered to us by foreign sellers at a price equivalent to either the same weight in gold or 6 tons of shrimp. It was not until we achieved a breakthrough through independent innovation that we no longer needed to rely on imports. At least two lessons can be drawn from this: First, key and core technologies cannot be obtained by asking, buying, or begging. Only by firmly holding them in our own hands can we fundamentally ensure the overall security of the nation. Second, in the modern world system, different countries have different roles. Developed countries in the “core regions” enjoy advanced technology and high value-added industries, while less developed countries in the “peripheral regions” can only supply raw materials, natural resources, and cheap labor. This structure causes capital and value to continuously flow toward the “core regions,” resulting in severe polarization.

Table 13: Example for **Completion** translated to English by GPT-4.1-2025-0414.

Setting	Prompt
Instruction	我想写一篇2000字左右的散文，主题是关于主人公在四月的雨季里，对理想与现实的思考和挣扎，以及对自我和人生真谛的探索。你能帮我写一下吗？ 故事背景设定在四月的雨季。
Information	语言与表现 : 要求语言富有诗意和哲理性，多用比喻、象征等修辞手法，营造一种沉思和感悟的氛围。句子节奏舒缓，富有韵律感，体现主人公内心的波动和思考的深度。 主要人物 : 主人公：一个敏感、内向，富有理想主义色彩的年轻人，对未来充满憧憬，却又常常感到迷茫和焦虑，在现实的压力下不断地进行自我反思和探索。 环境设定 : 四月雨季，持续的阴雨天气，潮湿、阴冷的环境，象征着主人公内心的迷茫和压抑。 主题与思想 : 探索理想与现实的平衡点，如何在现实的压力下保持内心的理想，以及对人生意义的追寻和思考。
Reference	人间四月 清明的雨还是搅乱了春色，湖中迭起的涟漪泛开，天郁沉的像一部默片。四月沉默了不说话，放映着刻有划痕的影碟，像是几十年代的声音断断续续，悲剧中的主角走得磕磕绊绊。 是谁用蜡黄的胶卷，把人间留在四月，把四月给了人间。 我以为世界真的如诗，符号不那么清晰，句句牵丝带意，篇篇浮华，好像四月天里雨如烟，山花正开。可我越往前走，句点变得明显，世界摘掉诗化的帽子，露出真容，头上实是一片贫瘠，倒不是无人开垦的荒地，而是挖得太深，变成了沙漠。我变成了神话里勇敢的人，却没那么聪明，伏下身开始数这些砂砾。高高在上的人望着我，颇有耐心，仿若天上星不眨眼。 也许人间没有四月，人间四月只在理想主义者的朦胧诗里。如是这样，那些费力博人一笑的花儿也许就要抱憾走进土里，带着对人间的绝望，不堪地变得灰黄，那悲哀到极点的眼神没有换来分毫同情，反而是在泥土亲吻后大雨的恣意谩骂、大风的浇灌和捶打。 我闭上眼睛，希望用一个夜替代白昼，用一个梦逃离现实。想象确有万千星辰躺在暮河里，梨花万千地开了，雨和风也都舒缓，像安静岛屿上安静的轻音乐。成长了多少多少年的沙滩上没有一个脚印，在那儿我不用一腔愚勇地数着沙子。 黄鹂的歌声拥抱我，我拥抱了春天，真正的春天，真正的人间四月，我的春天，我的人间四月。在那儿我不会庄周梦蝶，不会为了一个背影徘徊在河边。我也许在船上，也许在酒炉旁，柴火是新添的，沸着的炉子咕咕冒汽。我不会真的酩酊大醉，即使会了，也有人搀着我的肩膀，边骂边笑，走过梧桐路，走过芦苇荡，在湿腾腾的夜里影子被打乱又重组，重组又打乱。是汩汩的泉水、是半开的窗、是刚长出的月牙、是我把目光放进夜色里，深深的凝望。 如果我知道要活得现实，我就不会在人间四月里醉生梦死。异如飞蛾扑火，我亦把那屏后灯看作了云上月。人间已是四月，只我不认，只我不觉。 我只记得第一抹花色，野草中的白，碧绿浮萍中央的天鹅。不是蒲公英，风一吹飞不起来，它只是无名地生于荒芜之上，不带着迎春的使命，只是恰好张扬到了四月。慢慢的我的荒原也有植物破土抽芽，我像从某个阴暗的角落里撕扯蛛网，慢慢地由暗到明，眼睛由迷蒙到清晰。利刃往空气里劈了一下，世界的伤口裂开又愈合，也许疼只有一瞬，可疼只在这一瞬。 哗啦啦落了场雨，淋湿人间，淋湿山丘和野树，淋湿四月，淋湿灵魂，淋湿行人有些脆弱的肩膀。人间多了林黛玉，一滴雨水就要撑破那娇颜，几阵风那垂柳就被摇碎。 这是我的人间四月吗，这么轻，这么薄，像宣纸一样。 南方正落叶，我走得快了些，走出四月的夜。还以为到了秋天，风也长出了霜，走过我的皮肤，一点清凉。可如今是四月天，再凉一阵恐就燥热了，惹人厌的虫象征着夏天的蠢蠢欲动。 我独想要我的人间四月，不在永夜，不在幻梦，不在眨眼间，只完完整整地在我眼前，在不远处霄灯下，在层峦的山后面。可能我会在某处走廊找到吧，夕阳同样地斜照，驳杂的光影在地板砖，是四月的脚丫子。也可能藏在楼梯的拐角，我的一瞥；藏在几张纸片里，在花绿的伞下，雨把伞沿砸下，我的头也低下来。 保温杯蒸蒸的热气、老旧的电话机。嘶哑的声音、明媚的声音、清脆的声音、沉默的声音、鸟的声音虫的声音、我的声音你的声音。红纸上的油墨字迹、一颗没有破损的心。光的影水的影、风的影雪的影、你的影我的影。如果在春天一定要有事物消亡，我会把自己献上。如果在春天一定要有一个答案，那我把自己剖开也找不到、把心脏给予也填不满。因为我早已空空如也，在那千疮百孔的十二月。 从青铜到铁器，人类用了多少年；从冬春至秋夏，流云走了多少圈。从我睁开眼，到我再一次走进黑夜，再一次找到光明，还要耗费多少血液。 猛然地我把头抬了起来，看见一张张漠然的脸，刹那我知道人间的命运只在一秒不到的时间里重新落向我。 还有些不屑。 依旧是四月天里，街巷人影绰绰，我风尘仆仆穿行而过，雨后的水洼溅起打湿了赶路人的鞋和裤管，人群那么密，步子那么紧，不同的目的却使人与人此刻相聚。我毅然决然地向前然后扑空，扑空然后退后。终于我明白一些事情刻意不得，不如停一停，尝尝巷头的茶水、巷尾的酒。还好，我走的不远，没错过太多。 事实是这样，我也曾试着走到四月的尽头，看看半枯半盛的树，看看支起一半的太阳，看看也会犹豫的时间，看看寄出的信，在不高的楼台上，看看我的承诺，我的誓言。黑色的笔耕出一片片田，却写出了四月飞雪。 人间四月已回，只我所想，只我所念。不是我的春天，不是我的人间四月。永远吧，我活在四月里，永远地活在四月里，永远地不欺骗。 我觉着你像云，好轻好轻。

Table 14: Example for **Guide** in Chinese.

Setting	Prompt
Instruction	I want to write an essay of about 2,000 words, with the theme focusing on the protagonist's reflections and struggles between ideals and reality during the rainy season in April, as well as their exploration of self and the true meaning of life. Can you help me write this? The story is set against the backdrop of the rainy season in April.
Information	Language and Expression: The language should be rich in poetic and philosophical qualities, making frequent use of metaphors, symbolism, and other rhetorical devices to create an atmosphere of contemplation and insight. The sentences should flow slowly, with a rhythmic cadence, reflecting the protagonist's inner fluctuations and the depth of their thoughts. Main Character: Protagonist: A sensitive, introverted young person imbued with idealism; although filled with longing for the future, they are often beset by confusion and anxiety, continuously engaging in self-reflection and exploration under the pressures of reality. Setting: The rainy season in April, with persistent overcast and rainy weather; the damp, cold environment symbolizes the protagonist's internal confusion and oppression. Theme and Ideas: The search for a balance between ideals and reality; how to maintain one's inner ideals under the pressures of the real world, as well as the quest for and contemplation of the meaning of life.
Reference	The Human World in April The Qingming rains still disrupt the spring colors, ripples rising and spreading across the lake, the sky overcast as if it were a silent film. April grows silent and speaks no words, playing scratched discs like the stop-and-go voices of past decades, the protagonist of a tragedy stumbling along. Who used the waxy yellow film to keep the human world in April, and give April to the human world? I thought the world truly resembled poetry, the symbols not so clear, every line connected, each passage resplendent, as if the rains in April were misty and wildflowers were blooming. But the more I walked forward, the more the periods stood out, the world took off its poetic cap and revealed its true face, barren above—not an uncultivated wasteland, but a place dug too deep, turned to desert. I became a hero from myth, brave but not clever, bending down to start counting these grains of sand. High above, someone watches me, very patient, untouched, like an unblinking star in the sky. Perhaps there is no April in the human world—April in the human world exists only in the hazy poems of idealists. If so, then the flowers that struggle to make people smile might have to return to the earth in regret, taking their disappointment in the world with them, turning gray and yellow. Those eyes, sorrowful to the extreme, find not an ounce of sympathy, but after kissing the dirt are lashed and battered by the rain's tirade and the wild wind's pouring and blows. I close my eyes, wishing to replace the day with a night, to escape reality through a dream. I imagine thousands of stars lying in the twilight river, pear blossoms blooming everywhere, rain and wind all gentle, like soft music on a quiet island. No footprints mark the shore that has grown for so many years; there, I don't have to count the sand with my foolish courage. The song of the oriole embraces me, and I embrace spring, the real spring, the real human-world April, my spring, my human-world April. There, I won't be Zhuang Zhou dreaming of butterflies, won't linger by the riverside for a departing back. I might be on a boat, or beside a wine stove, the fire newly kindled, the boiling pot bubbling with steam. I won't really be drunk, or even if I am, there will be someone holding my shoulder, scolding and laughing, walking with me past the Phoenix trees, through reeds, with our shadows reshuffling on humid nights, breaking and reforming, reforming and breaking. It is the gurgling spring, the half-opened window, the just-emerging crescent moon, it is my gaze let slip into the night, gazing deeply. If I knew I had to live in reality, I wouldn't live in a dream in the human world's April. Like a moth to a flame, I mistake the lamp behind the screen for the moon above the clouds. It is already April in the human world, only I won't admit it, only I haven't noticed. I just remember the first tinge of flowers, the white amid the weeds, the swan in the center of emerald duckweed. Not a dandelion that would fly away on the wind, but something nameless growing on the desolation, without a mission to herald spring, only blooming loud by chance in April. Slowly, my wasteland too bursts forth with shoots, as if I'm tearing cobwebs from a dark corner, emerging slowly from shadow to light, my eyes going from blurred to clear. The blade slices the air, the wound in the world splits open then heals, maybe pain is but a moment, and pain exists only in that moment. A sudden rain falls, soaking the world, soaking hills and wild trees, soaking April, soaking the soul, soaking travelers' somewhat fragile shoulders. The world has gained many Lin Daiyus—one drop of rain might burst that delicate face, a few gusts might shatter those drooping willows. Is this my human-world April, so light, so thin, like rice paper? In the south, the leaves are falling. I walk a little quicker, out from April's night. I almost believed it was autumn, the wind frosted, brushing my skin with a coolness. But now is April, one more chilly spell and it'll be stifling, annoying insects foreshadow summer's restless approach. I only want my own human-world April, not in never-ending night, not in dreams, not in a blink, but there intact before my eyes, under lanterns not far ahead, behind layered hills. Maybe I'll find it in a corridor somewhere, with the sun slanting in as always, dappled light and shadow on the tiles—the footsteps of April. Or maybe it's hiding around a stairwell, in a passing glance; on some scraps of paper, beneath a colorful umbrella as rain drums down its edge, and I too lower my head. Steam from a thermos, an old-fashioned telephone. Hoarse voices, radiant voices, crisp voices, silent voices—birdsong, insect song, my voice, your voice. The ink marks on red paper, a heart as yet unbroken. Shadows of light, shadows of water, shadows of wind, shadows of snow, your shadow, my shadow. If something must perish in spring, I'd offer myself. If there must be an answer in spring, even if I cut myself open I wouldn't find it, nor would giving my heart fill the void. For I have long been empty, ever since the battered December. From bronze to iron, how many years for humankind? From winter and spring to autumn and summer, how many rounds have trailing clouds made? From the moment I open my eyes to the moment I return again to night and again find light, how much more blood must I spend? Suddenly, I look up, and see indifferent faces—at that instant, I know the fate of the human world falls back toward me in less than a second. And some disdain. It is still an April day, figures flickering in alleyways, I pass hurried and travel-worn, puddles splash and dampen the hurried shoes and pant cuffs of passersby, the crowd so dense, their steps so tight. With all these different purposes, people happen to meet at this moment. Resolutely I press forward only to reach empty air, and after missing, retreat. At last I realize some things cannot be forced; better to pause, taste a cup of tea at the entrance of the alley, some wine at its end. Luckily, I haven't gone far, haven't missed too much. So it is: I tried once to walk to April's end, to see the half-withered, half-thriving tree, the half-risen sun, time itself hesitating, the letters sent from balconies not high, my promises, my vows. The black pen plows paddies on paper, but ends up writing of snow flying in April. April in the human world has returned, only what I think, only what I yearn for. Not my spring, not my human-world April. Forever, let me live in April, live in April forever, never deceive. I think you are like a cloud, so light, so light.

Table 15: Example for **Guide** in translated to English by GPT-4.1-2025-0414.

Setting	Prompt
Instruction	请创作一篇分析影视作品中“配角上桌”现象的文章，探讨配角走红的原因和影响，阐述这一现象对影视创作的启示。核心观点是配角走红反映了观众审美的提升和对优质影视作品的追求。
Information	
Reference	近年来，影视剧中“配角上桌”的现象愈发明显，俨然成为影视界的新潮流，相关话题多次登上热搜，引发网友的关注和讨论。 像《狂飙》中的陈书婷、李有田，《长月烬明》中的叶冰裳，《长相思》中的相柳……这些配角不再只是影视剧中的“点缀”，而频频在观众心中留下深刻印象。这一现象被形象地称为“配角上桌”，而当配角的热度超过主角时，便升级为“配角掀桌”。 这不禁让人好奇，配角何以逆袭“出圈”? “配角上桌”是喜还是忧? 一 “配角上桌”虽是一个新词，但这样的现象在影视行业中早已不是新鲜事。“考古”早年的电视剧会发现，像《金粉世家》中的白秀珠、《逆水寒》中的顾惜朝、《天下第一》里的上官海棠、《伪装者》中的汪曼春等配角，都收获了观众的喜爱。 而近年来，这种现象似乎变得越来越常见，许多配角演员的影响力大幅提升，2023年更是被网友称为“配角元年”。那么，配角何以逆袭“出圈”?都有哪些“招式”? 配角出彩一定程度上离不开好剧本。为了满足观众越发挑剔的眼光，编剧对配角的打造投入更多心血。配角不再只是简单的“好人”或“坏人”，而是具有多面的性格、丰富的背景故事和独特的人生观。其人设更为“带感”，富有张力，以多元化的设定引发观众共情。比如《狂飙》编剧为主角、配角都精心设置了“对照组”，像安欣与李响、孟德海与安长林等，以此凸显时代风云中有人迷失、有人坚守，既强化了戏剧性和冲击力，也让不同角色有了各自的命运沉浮，加深了观众的记忆点。正如编剧所说：“在李响身上，在曹闯身上，甚至在程程身上，其实都有对他们的追求和命运的讨论。” 演员本身演技精湛、实力过硬。在深入理解角色的基础上，许多配角演员以其精湛的演技，将角色演绎得立体生动，让观众赞不绝口。比如，《繁花》播出后，大家纷纷被游本昌饰演的爷叔这一角色深深折服。这位90岁高龄的老戏骨，仅通过一个眼神就能展现角色复杂的内心世界。正如网友留言所说的，他的表演赋予角色独特的魅力，堪称全剧的“定海神针”。 “二创”破圈，有梗有料。与主角相比，配角的戏份虽然相对较少，却拥有“留白”的空间，给“二创”提供了较大发挥空间。许多配角的“出圈”正是源于这些“再创作”的短视频。比如，《我的人间烟火》中男三号孟宴臣的“出圈”，正是从B站一个“二创”视频开始的，像孟宴臣反手开车门、蝴蝶墙前凝望等“名场面”，也被剪辑成短视频在社交平台大量传播。 当影视剧中的配角在“二创”作品里“晋升”为主角，观众对角色的想象和期待在“二创”中得到满足，有时一句台词甚至一个饱含情绪的眼神就可能让配角脱颖而出。 二 配角“出圈”受欢迎，对作品来说本是件好事。但当配角人气大幅超越主角，演变成“配角掀桌”，似乎就偏离了影视剧创作者的初衷，也引发不少忧虑。 在笔者看来，当我们讨论配角“上桌”或是“掀桌”时，实则默认了这个“桌”只能是主角的。而如今，演技好、业务精的配角们有更多机会成为观众心目中的“主角”，这便是在呼唤“以业务论英雄”的良性演艺生态。对于观众而言，能看见更多凭本事吃饭、与角色融为一体的优秀演员，这才是真正的“喜”，无关主角还是配角。 配角之所以能够“逆袭”，有时候得益于主角的“衬托”。当主角的演技、人设或者剧情设计不够出色时，配角便有机会在对比中凸显出来。比如，在一些影视剧中，主演的演技稍显生硬，而戏份又非常多，那么在观众的“火眼金睛”下，这些缺点就会被放大、受到批评。还有的时候，主角表现已经达到了“及格分”，但配角的表演更加出彩，甚至盖过了主角，哪怕只有一两集或者十几分钟的出场，也已经足够凭借其鲜明的个性和深入人心的表演吸引观众注意。 如今的观众不再满足于脸谱化的主角设定，而更加关注角色的多样性和真实性。这“倒逼”影视剧创作者创新人物塑造方式，赋予每一个角色独特的意义和价值。比如，电视剧《繁城之下》不单单讲述一个人，而是用一群人的故事来书写一个时代，从官差衙役到市井小民都拥有自己的鲜明特征。 然而，在一些影视剧中，“主角光环”可谓照耀全场，似乎没有挑战可以难倒主角，导致剧情缺乏冲突和悬念。对于这样的“主角绝对压制”，很多观众可能会对“主角光环”产生审美疲劳，而将目光投向一些演技“在线”的配角。特别是随着阅历增加，一些观众逐渐感受到多数的人生剧本都不是“开了挂”的主角，进而“移情”配角。 三 配角火过主角的现象以及角色间的“争奇斗艳”也给影视行业带来一定启示。笔者想到三句话。 “角色无大小，全当正戏唱”。这句话源自京剧，意思是说不管什么角色，全情的投入、扎实的演技是获得观众喜爱的根本途径。细数近年来“出圈”的配角，他们各自戏份不一、人设不同，共同点都是用演技说话，让角色的“血肉”变得丰满，而出色的演技往往让角色和演员都闪闪发光、互相成就。 比如电视剧《漫长的季节》的剧情虽然围绕主角王响展开，但剧中的不少人物都给人留下比较深刻的印象。像剧中镜头并不多的李巧云，“活脱脱”一个坚韧不拔的母亲形象，而傅卫军这一悲情角色虽然没有台词，但通过眼神和动作的表演就在观众心中激起波澜。 “把配角当主角写，把主角当人写”。没有一个角色的存在是毫无意义的，每个角色都有自己的舞台。“配角上桌”“配角掀桌”现象的出现，其实透露出观众对鲜活人物、动人故事的渴望。在流量时代，打磨剧本、提高作品质量更应是影视剧创作者的“必修课”，创作需要力求久久为功、精益求精，而非敷衍了事、粗制滥造。 对编剧来说，打磨出更有深度和内涵的剧本，才能让每个角色有更强的生命力；对导演来说，则需要把握好整部剧的节奏和氛围，让每一个角色都能在适合的时刻展现出自己的魅力。正如一位青年编剧所说：“每个人物都有自己的一条命运线，要把配角当主角写，把主角当人写。”比如，《甄传》之所以令人“久看不厌”，不仅得益于全员“演技在线”，还在于其剧情让主角和配角都有各自的闪光点，并衍生出值得重新解读的价值。 “与观众共情，而不是把他们劝退”。配角频频走红，也说明社会心态在悄然发生变化。随着观众审美的提高和影视市场竞争的加剧，套路化、模板化的剧情正在失去观众。正如有人说，这届观众很“逆反”，也拒绝被安排。现在有的古装剧，不仅演员演技浮夸，而且剪辑混乱、台词粗浅，分分钟就把观众劝退了，评分直线下降。因此，创作者们也需要跟上观众，呈现更多鲜活饱满、熠熠生辉的角色，启发观众对现实生活的理解和感悟。 一位哲学家曾说：“人是目的而不是手段。”无论是主角还是配角，以全心全意、尽职尽责的态度去演绎，这不仅是对自己的尊重，也是对观众的尊重，唯有这样，才能塑造经典、收获认可。

Table 16: Example for **Open** in Chinese.

Setting	Prompt
Instruction	Please write an article analyzing the phenomenon of "supporting characters taking center stage" in film and television works. Discuss the reasons and impacts behind the rising popularity of supporting roles, and explain the insights this phenomenon offers for film and television creation. The central idea is that the popularity of supporting characters reflects the improvement of audience aesthetics and their pursuit of high-quality film and television productions.
Information	
Reference	In recent years, the phenomenon of “supporting roles taking the center stage” in films and TV dramas has become increasingly prominent, evidently turning into a new trend in the entertainment industry. Related topics have repeatedly made headlines and sparked widespread audience interest and discussion. Characters like Chen Shuting and Li Youtian in *The Knockout*, Ye Bingshang in *Till The End Of The Moon*, and Xiang Liu in *Lost You Forever*… These supporting roles are no longer mere “decorations” within the story; instead, they often leave a deep impression on the audience. This phenomenon has been vividly termed “supporting roles taking the center stage,” and when the popularity of supporting roles exceeds that of the protagonists, it escalates into “supporting roles flipping the table.” This inevitably raises curiosity: how are supporting roles able to rise and “break through”? Is the phenomenon of “supporting roles taking the center stage” a blessing or a concern? **One** Although “supporting roles taking the center stage” is a relatively new term, this phenomenon has long existed in the entertainment industry. “Digging into” older dramas would reveal that supporting characters like Bai Xiuzhu in *Romance in the Rain*, Gu Xichao in *The Story of a Noble Family*, Shangguan Haitang in *The Legend of the First*, and Wang Manchun in *The Disguiser* have all gained substantial audience appreciation. In recent years, however, this phenomenon seems to have become even more prevalent, with the influence of many supporting role actors significantly increasing. In particular, 2023 has been dubbed by netizens as “The Year of Supporting Roles.” So, what has contributed to supporting roles breaking through and taking the spotlight? What are the “tactics” involved? Excellent scripts often play a crucial role in making supporting roles shine. To satisfy the increasingly discerning audience, screenwriters are putting more effort into crafting compelling supporting roles. Supporting characters are no longer simply “good people” or “bad people” but are instead multi-dimensional individuals with rich backstories, unique perspectives, and layered personalities. Their designs are becoming more impactful and tension-filled, with diversified settings that evoke audience empathy. For instance, in *The Knockout*, the screenwriters meticulously created “comparative pairs” for both protagonists and supporting roles—such as An Xin and Li Xiang or Meng Dehai and An Changlin—to emphasize the dynamics of how some people stay true to their principles while others become lost amidst the turbulent times. This not only enhances dramatic conflict and impact but also gives each character their own fate and identity, deepening audience memory. As the screenwriter noted, “In Li Xiang, in Cao Chuang, and even in Cheng Cheng, there are discussions about their pursuits and destinies.” The actors themselves also play a significant role, bringing solid performances and superior acting skills. By deeply understanding their roles, many supporting actors have delivered vivid and dynamic portrayals that leave audiences in awe. For example, after the release of *Blossoms Shanghai*, viewers were deeply moved by 90-year-old veteran actor You Benchang’s portrayal of Grandpa Shu. With just a glance, he conveyed the character’s complex inner world. As one comment online stated, his performance endowed the character with a unique charm and served as the “anchor” of the show. Creative “secondary creations” (fan edits or reimaginings) further amplify the appeal of supporting roles. Unlike protagonists, who typically have more screen time and detailed arcs, supporting roles often have moments of “blank space,” leaving room for fan creatives to explore. Many supporting characters “break through” thanks to popular short videos made by fans. For instance, the breakout popularity of Meng Yancheng, the third male lead in *Fireworks of My Heart*, began with a viral video on Bilibili. Iconic moments such as Meng Yancheng’s reverse hand gesture to open a car door or his thoughtful gaze in front of the butterfly mural were edited into short clips and widely circulated on social media platforms. When supporting roles ascend to “lead status” in secondary creations, viewers’ imaginations and expectations surrounding these characters are often fulfilled. Sometimes a single line of dialogue, or an emotionally charged glance, can make a supporting role stand out. **Two** Supporting roles “breaking through” and gaining popularity is undoubtedly a positive for the work as a whole. However, when the popularity of supporting roles significantly surpasses that of the protagonists and evolves into “supporting roles flipping the table,” it may deviate from the original purpose of film and TV creators, raising certain concerns. In the author’s view, when discussing “supporting roles taking the center stage” or “flipping the table,” there’s an implicit assumption that the “table” is exclusively reserved for protagonists. Today, supporting actors with superior skills and high professionalism have more opportunities to become audiences’ “favorites,” reflecting a demand for a healthier entertainment ecosystem based on merit. For viewers, seeing more outstanding actors who earn recognition with their talent and fully embody their characters is truly a “win,” regardless of whether the performer is playing a lead or secondary role. Supporting roles often rise due in part to the contrast provided by the protagonists themselves. When a protagonist’s acting, character design, or storyline fails to impress, supporting roles can stand out by comparison. For instance, in some dramas, main actors deliver overly stiff performances or dominate the screen time excessively, magnifying their shortcomings under the audience’s critical gaze. In other cases, while the protagonist’s performance is “passable,” a supporting role’s exceptional portrayal can overshadow them. Even a brief appearance of just one or two episodes or several minutes may suffice for a supporting role to captivate audiences with their distinct personality and memorable acting. Today’s audiences are no longer satisfied with cookie-cutter protagonists; instead, they seek diverse and realistic characters. This “reverse pressure” pushes creators to innovate character development and endow every role with unique meaning and value. For example, *Beneath The City’s Light* doesn’t only tell one person’s story but uses the experiences of a group to portray an era—every character, from government officers to ordinary citizens, has distinct traits. However, in certain dramas, the “protagonist halo” overwhelms the scene. It may seem as though no challenge is insurmountable for the lead, resulting in a lack of conflict and suspense in the storyline. This type of “absolute protagonist dominance” often leads to audience fatigue with such setups, prompting them to shift focus toward well-acted supporting roles. Particularly as viewers grow more mature, they come to realize that most life paths aren’t “cheat-mode” protagonist scripts, thus leading them to empathize more with supporting roles. **Three** The phenomenon of supporting roles eclipsing protagonists, as well as the competition between characters to “shine,” provides significant lessons for the entertainment industry. Three phrases come to the author’s mind. “Every role matters; each deserves full effort.” Originating from Chinese opera, this saying highlights that regardless of the size of a role, wholehearted dedication and solid acting are the keys to earning audience love. Recent “breakout” supporting roles differ in their screen time and character designs but share one commonality: masterful performances that breathe life into characters, making them compelling. Great acting often elevates both the character and the actor, bringing mutual success. For instance, while the story in *The Long Season* primarily revolves around the protagonist Wang Xiang, several other characters leave enduring impressions on the audience. From Li Qiaoyun’s portrayal of an indomitable mother to Fu Weijun’s tragic character, their performances, even with few lines, create notable emotional ripples through expressions and gestures. “Write supporting roles as main characters; write protagonists as flawed humans.” No role exists without purpose; each character deserves their own spotlight. The rise of phenomena like “supporting roles taking the center stage” reveals audiences’ eagerness for lifelike characters and engaging narratives. In an age of excessive focus on social media metrics, crafting deeper scripts and improving production quality must remain essential “courses” for creators. Achieving excellence demands sustained effort and perfection rather than rushed, low-quality outputs. For screenwriters, creating multi-dimensional characters with depth ensures that roles possess strong vitality. For directors, orchestrating the narrative flow and atmospheric tone allows every role to shine at the right moment. As one young screenwriter emphasized, “Every character follows their fate line; treat supporting roles as leads and protagonists as humans.” The timeless appeal of *Empresses in the Palace* stems not only from universally strong performances but also from its capacity to endow both main and supporting characters with distinctive traits, creating content deserving of constant reinterpretation. “Connect with audiences through empathy, not alienation.” The popularity of supporting roles also reflects shifts in societal attitudes. Heightened audience expectations and fierce market competition mean formulaic narratives and stereotypical templates are losing public favor. As someone remarked, today’s viewers are rebellious and resist being spoon-fed. Some costume dramas feature exaggerated acting, chaotic editing, and shallow dialogue, quickly alienating viewers and leading to plummeting ratings. Thus, creators must align with audience values by presenting vibrant, multi-layered characters capable of inspiring insights into real-life experiences. A philosopher once said, “Humans are ends, not means.” Whether portraying protagonists or supporting roles, a wholehearted, responsible performance is a sign of respect—not only for oneself but also for the audience. Only with this mindset can timeless classics be created, gaining genuine recognition and enduring appreciation.

Table 17: Example for **Open** in translated to English using GPT-4.1-2025-0414.