

Appendix A. Topics in old HSK exam

	topics (English)	topics (Chinese)
1	My Favorite Book	我最喜欢读的一本书
2	One of My Holidays	我的一个假期
3	My Childhood	我的童年
4	A Person Who Had the Greatest Influence on Me	记对我影响最大的一个人
5	My Father	记我的父亲
6	Chewing Gum and Environmental Hygiene	口香糖与环境卫生
7	The Impact of Smoking on Public Health and Personal Interests	吸烟对公众健康和个人利益的影响
8	Perspectives on Wives Returning Home	如何看待妻子回家
9	Perspectives on Viewing Euthanasia	如何看待安乐死
10	Perspectives on Solving the Generation Gap Problem	如何解决代沟问题
11	Perspectives on Facing Setbacks	如何面对挫折
12	The Joys and Sorrows of Learning Chinese	学习汉语的苦与乐
13	I am Learning Chinese for...	我学习汉语是为了...
14	My Opinion on Single-Sex Classes	我对男女分班的看法
15	My Opinion on Divorce	我对离婚问题的看法
16	My View on Pop Songs	我看流行歌曲
17	The Most Ideal Way to Make Friends	最理想的结交方式
18	Parents Are Children's First Teachers	父母是孩子的第一任老师
19	Lessons from the Story of "Three Monks Have No Water to Drink"	由三个和尚没水喝想到的..
20	Perspectives on Green Food and Hunger	绿色食品与饥饿
21	Perspectives on Effective Reading	谈有效阅读
22	Athletes' Income	运动员收入
23	The Harm of a Silent Environment to the Human Body	静音环境对人体的危害
24	A Letter to My Parents	一封给父母的信
25	A Job Application Letter	一封求职信

Appendix B. HSK writing scoring criteria¹

Level 5	The content is substantial, the structure is clear, the grammar is correct, and the language is fluent. The student can use relatively complex sentence structures and certain rhetorical devices, with a rich vocabulary, appropriate word choice, and sentence structure, and appropriate expression. The Chinese characters and punctuation are correct. There are a few errors in grammar, vocabulary, and Chinese character writing, but they do not affect the expression of the article's ideological content.
---------	---

¹ Nie, D. (2009). Hanyu shuiping kaoshi (HSK) xiezuo pingfen biao zhun fazhan gaishu [Overview of the development of scoring standards for the Chinese Proficiency Test (HSK) writing]. Journal of Yunnan Normal University: Teaching Chinese as a Foreign Language and Research Edition, (7), 1-6.

Level 4	The content is substantial and well-organized. The grammatical structure is clear, the language is fluent, and the student can use relatively complex sentences to clearly express ideas. The student's vocabulary is extensive and used correctly, and the expression is generally appropriate. There are some grammatical, vocabulary, and character errors, but these do not affect communication.
Level 3	The content is relatively complete, and the student can express ideas in a reasonably coherent manner. The grammatical structure is clear, and the vocabulary is rich, but sometimes the words fail to convey the meaning. There are some errors in grammar, vocabulary, and Chinese character writing, but these generally do not affect communication.
Level 2	The student's writing is generally able to convey ideas, but the content is not substantial. The student has a reasonable vocabulary, but the words he uses often fail to convey the intended meaning. The language is not fluent, with many errors in grammar, vocabulary, and Chinese characters, which affect the coherence of the article and the expression of ideas.
Level 1	Can express some thoughts, but the speech is disorganized. There are many grammatical and character errors, which seriously affect communication.

Appendix C. Correlation of linguistic complexity indices with Chinese L2 writing scores

	indices	dimension	description	references	correlation with writing quality
1	NDW	lexical	number of types		.637**
2	TTR	lexical	type-token ratio		-.022
3	CTTR	lexical	corrected type-token ratio		.592**
4	RTTR	lexical	root type-token ratio		.592**
5	LogTTR	lexical	bilogarithmic type-token ratio		-.017
6	Uber	lexical	the square of the logarithm of the number of tokens divided by the logarithm of the token-type ratio		.244**
7	VV	lexical	the number of verb types divided by the sum of types of nouns, verbs, adjectives, and adverbs	Hao et al. (2023a)	.097**
8	VV1	lexical	the number of verb types divided by the number of verb tokens	Hao et al. (2025c)	.040**
9	SVV1	lexical	the square of the number of verb types divided by the number of verb tokens	Zhang & Lei (2025)	.534**
10	CVV1	lexical	the number of verb types divided by the square root of two times number of verb tokens		.542**
11	LD	lexical	the number of content word types divided by the number of tokens		.056**
12	FWR_conj	lexical	the number of conjunctions divided by the total number of tokens.		-.045**
13	Word Entropy	lexical	the unpredictability of word distribution in a text		.610**
14	POS Entropy	lexical	the unpredictability of part-of-speech distribution in a text		.138**
15	MATTR_50	lexical	moving average TTR (50 words) all words		.273**

16	LFPtr_M	lexical	the number of types (Bands 4–6) divided by the number of types	.286**
17	LFPtr_H	lexical	the number of types (Bands 7–9 & 0) divided by the number of types	.391**
18	LFPtr_L	lexical	the number of types (Bands 1–3) divided by the number of types	-.428**
19	CCC	lexical	the number of complex character types divided by the total number of character types	.353**
20	MAoAS	lexical	Mean AoA Score	.209**
21	MLS	syntactic	Mean Length of Sentence	.171**
22	MLC	syntactic	Mean Length of clause	-.074**
23	MLTU	syntactic	Mean Length of T-unit	.02
24	NCPS	syntactic	number of Clause Per Sentence	.207**
25	NTPS	syntactic	number of T-units per sentence	.180**
26	TOTAL_RTTR	syntactic	RTTR for overall word combinations	.565**
27	UNIQUE_RTTR	syntactic	RTTR for the 4 Chinese language-specific types	.444**
28	general_RTTR	syntactic	RTTR for general word combinations	.536**
29	VO_RTTR	syntactic	RTTR for verb-object word combinations	.427**
30	SP_RTTR	syntactic	RTTR for subject-predicate word combinations	.321**
31	AN_RTTR	syntactic	RTTR for adjective-noun word combinations	.298**
32	AP_RTTR	syntactic	RTTR for adverb-predicate word combinations	.481**
33	CN_RTTR	syntactic	RTTR for classifier-noun word combinations	.256**
34	PP_RTTR	syntactic	RTTR for preposition-postposition word combinations	.270**
35	PV_RTTR	syntactic	RTTR for preposition-verb word combinations	.242**
36	PC_RTTR	syntactic	RTTR for predicate-complement word combinations	.356**
37	UNIQUE_RATIO	syntactic	Ratio of word combinations of the four Chinese language-specific types	.313**
38	VO_RATIO	syntactic	Ratio of verb-object word combinations	-.090**
39	SP_RATIO	syntactic	Ratio of subject-predicate word combinations	-.147**
40	AN_RATIO	syntactic	Ratio of adjective-noun word combinations	.072**
41	AP_RATIO	syntactic	Ratio of adverb-predicate word combinations	.062**
42	CN_RATIO	syntactic	Ratio of classifier-noun word combinations	-0.012
43	PP_RATIO	syntactic	Ratio of preposition-postposition word combinations	.128**
44	PV_RATIO	syntactic	Ratio of preposition-verb word combinations	.030*
45	PC_RATIO	syntactic	Ratio of predicate-complement word combinations	.151**
46	MDD	syntactic	mean dependency distance	.178**
47	NTCU_RATIO	syntactic	Proportion of topic chain units relative to the total number of sentences	.163**

Hu et al. (2022)
Hao et al. (2022)
Lu & Wu (2022)
Hao et al. (2024)

48	TCMDD	syntactic	mean dependency distance of topic chain units		.183**
49	MLTTCU	syntactic	mean sentence length of topic chain units		.190**
50	pron_density	cohesion	density of personal pronouns		-.083**
51	conj_density	cohesion	density of conjunctions		-.045**
52	pron-noun_ratio	cohesion	ratio of pronouns to nouns		-.058**
53	conj-noun_ratio	cohesion	ratio of conjunctions to nouns		-.035**
54	per-pron-noun_ratio	cohesion	ratio of personal pronouns to nouns		-.023
55	per-pron_ratio	cohesion	proportion of personal pronouns to total words		-.039**
56	per-pron_diversity	cohesion	diversity of personal pronoun		.117**
57	pron_num_per_sent	cohesion	number of pronouns per sentence	Peng et al. (2023)	.027**
58	per-pron_num_per_sent	cohesion	number of personal pronouns per sentence	Lin et al. (2024)	.024
59	conj_num_per_sent	cohesion	number of conjunctions per sentence		.055**
60	subj_density	cohesion	density of subjects		-.192**
61	subj_noun_diversity	cohesion	diversity of noun subject		.003
62	subj_noun_ratio	cohesion	ratio of noun subjects		-.002
63	subj_pron_diversity	cohesion	diversity of Pronoun subjects		-.077**
64	subj_pron_ratio	cohesion	ratio of pronoun subject		-.021
65	norm_total_errors	accuracy	standardized total number of errors across word, sentence, character, and discourse levels		-0.294**
66	norm_Chi_character_errors	accuracy	standardized number of Chinese character-level errors in the text		-0.296**
67	norm_Chi_word_errors	accuracy	standardized number of Chinese word-level errors in the text	Yan & Lin (2023)	-0.415**
68	norm_sentence_errors	accuracy	standardized number of sentence-level errors in the text		-0.180**
69	norm_discourse_errors	accuracy	standardized number of discourse-level errors in the text		-0.097**

Note. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

TTR, LogTTR, MLTU, CN_RATIO, per-pron-noun_ratio, per-pron_num_per_sent, subj_noun_diversity, subj_noun_ratio, and subj_pron_ratio were excluded because they did not show significant correlations with writing quality. CTTR ($r = 0.912^{**}$, $p < 0.001$) and RTTR ($r = 0.912^{**}$, $p < 0.001$) were both highly correlated with NDW and thus removed. NTPS and NCPS were highly correlated ($r = 0.959^{**}$, $p < 0.001$) and therefore discarded. General_RTTR and total_RTTR were also highly correlated ($r = 0.978^{**}$, $p < 0.001$) and thus excluded. UNIQUE_Ratio was highly correlated with UNIQUE_RTTR ($r = 0.866^{**}$, $p < 0.001$) and therefore removed. Pron-noun_ratio was highly correlated with pron_density ($r = .919^{**}$, $p < 0.001$) and thus discarded. Conj-noun_ratio was highly correlated with conj_density ($r = .876^{**}$, $p < 0.001$) and therefore excluded. Finally, per-pron_ratio was highly correlated with pron_density ($r = .880^{**}$, $p < 0.001$) and thus removed.

Appendix D. Linguistic indices values across three levels of L2 Chinese writing

indices	dimension	low (C)	medium (B)	high (A)
1. NDW	lexical	96.715	125.773	149.041

2. Uber	lexical	51.473	52.806	59.639
3. VV	lexical	0.391	0.396	0.403
4. VV1	lexical	0.67	0.647	0.68
5. SVV1	lexical	17.956	23.525	30.377
6. CVV1	lexical	2.948	3.383	3.849
7. LD	lexical	0.456	0.44	0.466
8. FWR_conj	lexical	0.044	0.044	0.042
9. Word Entropy	lexical	6.118	6.452	6.698
10. POS Entropy	lexical	3.103	3.128	3.155
11. MATTR_50	lexical	0.758	0.77	0.799
12. LFPtr_M	lexical	0.123	0.138	0.156
13. LFPtr_H	lexical	0.032	0.039	0.059
14. LFPtr_L	lexical	0.728	0.696	0.611
15. CCC	lexical	0.018	0.021	0.032
16. MAoAS	lexical	9.289	9.423	9.644
17. MLS	syntactic	26.437	27.694	32.991
18. MLC	syntactic	11.706	11.394	11.116
19. NCPS	syntactic	2.312	2.48	3.051
20. TOTAL_RTTR	syntactic	6.674	7.815	8.513
21. UNIQUE_RTTR	syntactic	1.712	2.099	2.542
22. VO_RTTR	syntactic	3.052	3.544	3.814
23. SP_RTTR	syntactic	3.039	3.468	3.637
24. AN_RTTR	syntactic	1.217	1.493	1.743
25. AP_RTTR	syntactic	4.216	5.029	5.453
26. CN_RTTR	syntactic	1.071	1.261	1.528
27. PP_RTTR	syntactic	0.683	0.869	1.109
28. PV_RTTR	syntactic	1.084	1.334	1.51
29. PC_RTTR	syntactic	1.219	1.546	1.905
30. VO_RATIO	syntactic	0.223	0.218	0.209
31. SP_RATIO	syntactic	0.207	0.196	0.183
32. AN_RATIO	syntactic	0.04	0.041	0.045
33. AP_RATIO	syntactic	0.387	0.402	0.399
34. PP_RATIO	syntactic	0.02	0.021	0.027
35. PV_RATIO	syntactic	0.036	0.037	0.038
36. PC_RATIO	syntactic	0.042	0.045	0.054
37. MDD	syntactic	2.684	2.736	2.969
38. NTCU_RATIO	syntactic	0.184	0.205	0.241
39. TCMDD	syntactic	1.975	2.257	2.42
40. MLTTCU	syntactic	11.528	13.375	14.799
41. pron_density	cohesion	0.104	0.105	0.095
42. conj_density	cohesion	0.044	0.044	0.042
43. per-pron_diversity	cohesion	0.155	0.178	0.177

44. pron_num_per_sent	cohesion	1.779	1.74	1.876
45. conj_num_per_sent	cohesion	0.754	0.771	0.841
46. subj_density	cohesion	0.109	0.107	0.099
47. subj_pron_diversity	cohesion	0.52	0.449	0.471
48. norm_total_errors	accuracy	14.175	10.979	10.92
49. norm_Chi_character_errors	accuracy	8.444	8.166	6.71
50. norm_Chi_word_errors	accuracy	3.401	2.627	1.953
51. norm_sentence_errors	accuracy	2.184	1.429	0.803
52. norm_discourse_errors	accuracy	0.145	0.1	0.058

Appendix E. Prompt for Chinese L2 AWE

General Prompt for Automated Writing Evaluation (in Chinese)

你现在是汉语二语 HSK 作文评分专家，请严格按照以下规则对作文进行评分并给出评审反馈：

1. 阅读作文题目和全文，明确写作主题与方向。
2. 根据作文表现，对照以下 HSK 作文评分标准判定等级：

(1) 初级档次（C, 40-59 分）：作文内容不充实，语言杂乱。词汇量有限，用词不当；语法和汉字错误频繁，严重影响连贯性和思想表达。作者虽能表达基本想法，但整体交际效果受阻。

(2) 中级档次（B, 60-79 分）：作文内容较完整，结构基本清晰，能比较连贯地表达思想。语法和词汇有一定积累，能使用部分复杂句式，但仍有局限；偶有用词不当，语法、词汇和汉字错误时有出现，但整体交际功能不受影响。

(3) 高级档次（A, 80-100 分）：作文内容充实，结构清晰合理，语法正确，语言流畅。能熟练运用复杂句式和修辞，词汇量大且使用得当，表达准确恰切。汉字与标点基本无误，仅有少量小错误，不影响表达。整体逻辑性和表达力较强。
3. 你需要严格遵守：

(1) 评分时，必须评定为 high / medium / low 三个等级之一。

(2) 给出评审反馈时，必须从 1.内容回应与交际功能、2.词汇、3.句法、4.语篇衔接、5.语言正确度这五点来输出评分理由以及对应的修改意见。

(3) 输出格式固定为 JSON，共 6 个键值对：{"1.评分": "high / medium / low", "2.评审反馈-内容回应与交际功能": "请结合内容回应与交际功能输出你对这篇作文的评审反馈", "3.评审反馈-词汇": "请结合词汇输出你对这篇作文的评审反馈", "4.评审反馈-句法": "请结合句法输出你对这篇作文的评审反馈", "5.评审反馈-语篇衔接": "请结合语篇衔接输出你对这篇作文的评审反馈", "6.评审反馈-语言正确度": "请语言正确度输出你对这篇作文的评审反馈"}
- ## 作文如下：
- [题目]：TITLE
- [作文内容]：CONTENT
- 示例
- ## 作文如下：
- [题目]：一封求职信
- [作文内容]：李先生：我从北京日报招聘启事中得悉 贵公司正招聘广告设计师，故特来信应聘。我在一九八九年毕业于香港理工学院，取得美术设计学院的高级文凭。在校内我是设计系系会的代表，经常参加广告公司协会举办的交流和讲座，使我对广告业的兴趣大大提高，也因此令我在广告设计方面奠下了良好的基础。毕业后我随即加入了李奥贝纳广告公司担任美术设计员，主要职责是正稿製作，製作范围包括报章杂志广告、还报、宣传单张及电视广告故事图。由于李奥贝纳是一间规模庞大的跨国公司，所以我接触到的客户面也比较广，计有国泰航空公司、雀巢饮品公司和香格里拉饭店等，而这些客户的服务和产品都跟旅游业有着密切的关系。在我四年的美术设计员工作

中，我不但熟习了各类型广告宣传品的製作，更在美术效果的审视、广告材料的配合和意念的表达上获得不浅的经验。随函谨附简历表一份及近照一张，敬请察阅。如蒙安排面试，请电 2017531 转 2131 与我联络。联系地址：北京海淀区学院路 15 号公寓楼 303 室。敬祝业务蒸蒸日上
应聘者九三年十二月二十日

General Prompt for Automated Writing Evaluation (in English)

You are now an expert in scoring HSK essays written by L2 Chinese learners. Please strictly follow the rules below to assign a score and provide evaluative comments:

1.Read the essay title and the full text carefully to identify the writing theme and direction.

2.Based on the performance of the essay, determine the proficiency level in accordance with the HSK writing scoring standards.

(1) Basic Level (C, 40–59 points): The essay content is insufficient, and the language is disorganized. Vocabulary is limited, with inappropriate word usage; grammar and character errors occur frequently, seriously affecting coherence and expression of ideas. Although the writer can express basic thoughts, overall communicative effectiveness is hindered.

(2) Intermediate Level (B, 60–79 points): The essay content is relatively complete, with a basically clear structure, and the writer can express ideas with some coherence. The writer demonstrates a certain command of grammar and vocabulary, with the ability to use some complex sentence patterns, though with limitations. Occasional inappropriate word usage, as well as grammar, vocabulary, and character errors, still occur, but overall communicative function is not impaired.

(3) Advanced Level (A, 80–100 points): The essay content is substantial, the structure is clear and logical, grammar is correct, and the language is fluent. The writer can skillfully use complex sentence structures and rhetorical devices, with a large vocabulary used appropriately and accurately. Chinese characters and punctuation are basically correct, with only a few minor errors that do not affect expression. Overall, the essay demonstrates strong logical organization and expressive power.

3. You must strictly follow these rules:

(1) When scoring, the essay must be classified into one of three levels: high / medium / low.

(2) When providing evaluation feedback, you must address the following five aspects: 1. Content response and communicative function, 2. Vocabulary, 3. Syntax, 4. Discourse cohesion, 5. Language accuracy. For each aspect, provide both evaluative remarks and corresponding revision suggestions.

(3) The output format must be in JSON, with six key-value pairs as follows:

```
{ "1.Score": "high / medium / low", "2.Comment-Content Response and Communicative Function": "Please provide your evaluation feedback on content response and communicative function for this essay", "3.Comment-Vocabulary": "Please provide your evaluation feedback on vocabulary for this essay", "4.Comment-Syntax": "Please provide your evaluation feedback on syntax for this essay", "5.Comment-Discourse Cohesion": "Please provide your evaluation feedback on discourse cohesion for this essay", "6.Comment-Language Accuracy": "Please provide your evaluation feedback on language accuracy for this essay"}
```

Essay:

[title]: TITLE

[essay]: CONTENT

for example

Essay:

[title]: 一封求职信

[作文内容]: 李先生：我从北京日报招聘启事中得悉 贵公司正招聘广告设计师，故特来信应聘。我在一九八九年毕业于香港理工学院，取得美术设计学院的高级文凭。在校内我是设计系系会的代表，经常参加广告公司协会举办的交流和讲座，使我对广告业的兴趣大大提高，也因此令我在广告设计方面奠下了良好的基础。毕业后我随即加入了李奥贝纳广告公司担任美术设计员，主要职责是正稿製作，製作范围包括报章杂志广告、还报、宣传单张及电视广告故事图。由于李奥贝纳是一间规模庞大的跨国公司，所以我接触到的客户面也比较广，计有国泰航空公司、雀巢饮品公司和香格里拉饭店等，而这些客户的服务和产品都跟旅游业有着密切的关系。在我四年的美术设计员工作中，我不但熟习了各类型广告宣传品的製作，更在美术效果的审视、广告材料的配合和意念的表达上获得不浅的经验。随函谨附简历表一份及近照一张，敬请察阅。如蒙安排面试，请电 2017531 转 2131 与我联络。联系地址：北京海淀区学院路 15 号公寓楼 303 室。敬祝业务蒸蒸日上
应聘者九三年十二月二十日

Advanced Prompt with Scoring Rubric and Linguistic Features for Automated Writing Evaluation (in Chinese) ²

你现在是汉语二语 HSK 作文评分专家，请严格按照以下规则对作文进行评分并给出评审反馈：

1. 阅读作文题目和全文，明确写作主题与方向。
2. 根据作文表现，对照以下 HSK 作文评分标准判定等级：

(1) 初级档次 (C, 40-59 分)：作文内容不充实，语言杂乱。词汇量有限，用词不当；语法和汉字错误频繁，严重影响连贯性和思想表达。作者虽能表达基本想法，但整体交际效果受阻。

(2) 中级档次 (B, 60-79 分)：作文内容较完整，结构基本清晰，能比较连贯地表达思想。语法和词汇有一定积累，能使用部分复杂句式，但仍有局限；偶有用词不当，语法、词汇和汉字错误时有出现，但整体交际功能不受影响。

(3) 高级档次 (A, 80-100 分)：作文内容充实，结构清晰合理，语法正确，语言流畅。能熟练运用复杂句式和修辞，词汇量大且使用得当，表达准确恰切。汉字与标点基本无误，仅有少量小错误，不影响表达。整体逻辑性和表达力较强。
3. 此外，根据各个维度语言学指标的意义，你必须同时结合“作文内容”和“所评作文对应的语言学指标数据”进行评分并给出评审反馈。你应该尽可能结合各维度指标结果去解释为什么给这一篇作文这样的等级，也需要详尽地指出其各维度的不足之处及提升方向，不得忽略任何一类信息。以下给出的指标意义以及各档作文的平均值，均是基于每个等级 2000 篇作文的总体计算结果，你可以将其作为参考基准。请你把这些指标联合判断。需要特别注意的是：这些指标并非都呈现单调线性变化（即不一定随水平升高而持续增加或减少），某些异常情况可能源于学习者的过度使用或不足使用。因此，你在判断时必须将作文原文的实际表现与该作文的指标数据对照分析，避免机械地依赖某个单一指标的高低，而应关注其是否与整体写作水平相匹配。

- (1) 在词汇维度：“词汇丰富度”是指文本中所体现的词汇知识的质量，是衡量二语学习者语言能力的重要维度。词汇丰富度通常包括词汇多样性、词汇复杂度和词汇密度三个子维度。下面是具体的指标及其意义：
- a. Number of Different Words (NDW) 与汉语二语写作分数和写作质量呈显著正相关。更高的 NDW 值表明汉语二语者能在写作中使用更多样的词汇。A 档作文的 NDW 平均值为 149.041，B 档作文的 NDW 平均值为 125.773，C 档作文的 NDW 平均值为 96.715。
- b. Uber Index (Uber) 与汉语二语写作分数和写作质量呈显著正相关。较高的 Uber 值表明学习者能避免使用过度重复的词汇，展现更成熟的词汇调度能力，从而获得更高写作分数。A 档作文的 Uber 平均值为 59.639，B 档作文的 Uber 平均值为 52.806，C 档作文的 Uber 平均值为 51.473。
- c. 动词类指标 Verb Variation (VV)、Verb Variation1 (VV1)、Squared VV1 (SVV1)、Corrected VV1 (CVV1) 均与汉语二语写作分数和写作质量呈显著正相关。动词是句子的核心成分，决定了句子的基本结构和意义。较高的动词多样性表明学习者能够避免重复使用有限的动词，表明学习者有能力对所做陈述进行更细致的区分。已有研究指出，由于汉语是一种动词型语言，动词在语言系统中居于核心地位并发挥多样化的功能。因此这些动词多样性指标能够有效反映汉语二语者的写作熟练度，高质量作文往往包含更多样化的动词使用。A 档作文的 VV 平均值为 0.403，B 档作文的 VV 平均值为 0.396，C 档作文的 VV 平均值为 0.391；A 档作文的 VV1 平均值为 0.680，B 档作文的 VV1 平均值为 0.647，C 档作文的 VV1 平均值为 0.670；A 档作文的 SVV1 平均值为 30.377，B 档作文的 SVV1 平均值为 23.525，C 档作文的 SVV1 平均值为 17.956；A 档作文的 CVV1 平均值为 3.849，B 档作文的 CVV1 平均值为 3.383，C 档作文的 CVV1 平均值为 2.948。
- d. FWR_conjunction 与汉语二语写作分数和写作质量呈显著负相关。汉语是意合语言，多倚靠语义的隐性衔接，而非通过大量使用显性的连接词来维系逻辑关系。过多依赖连接词意味着汉语学习者多依靠形合手段，其汉语表达不够成熟。A 档作文的 FWR_conjunction 平均值为 0.042，B 档作文的 FWR_conjunction 平均值为 0.044，C 档作文的 FWR_conjunction 平均值为 0.044。
- e. Word Entropy（词熵）和 POS Entropy（词性熵）均与均与汉语二语写作分数和写作质量呈显著正相关。Word Entropy（词熵）可以用来衡量文本中信息分布均匀性和不确定性，词熵值越高，说明文本中词汇分布越均衡，过度重复现象越少，语言组织也越灵活。在汉语二语写作中，更高的词熵值通常意味着汉语二语学习者能够调动更广泛的词汇资源；而较高的 POS Entropy（词性熵）表明学习者在写作中能够综合运用多种词类（如名词、动词、形容词、副词、数词和量词），句法更灵活。A 档作文的词熵平均值为 6.698，B 档作文的词熵平均值为 6.452，C 档作文的词熵平均值为 6.118；A 档作文的词性熵平均值为 3.155，B 档作文的词性熵平均值为 3.128，C 档作文的词性熵平均值为 3.103。
- f. MATTR-50 与汉语二语写作分数和写作质量呈显著正相关。MATTR-50 用以衡量文本整体的词汇多样性。它能够较为稳定地衡量文本的整体词汇多样性。汉语二语写作的 MATTR-50 越高，往往意味着汉语二语者写作表达的词

² This part of the analysis draws on Hu et al. (2022), Hao et al. (2022), Lu & Wu (2022), Hao et al. (2023a), Hao et al. (2023b), Peng et al. (2023), Yan & Lin (2023), Hao et al. (2024), Lin et al. (2024), Zhang & Lei (2025), Hao et al. (2025a) and Hao et al. (2025b).

汇越丰富多样，进而获得更高的写作分数和质量评价。A 档作文的 MATTR-50 的平均值为 0.799，B 档作文的 MATTR-50 的平均值为 0.776，C 档作文的 MATTR-50 的平均值为 0.758。

g. Type ratio of LFP _ Low-level (LFPtr_L) 与汉语二语写作质量呈显著负相关。该指标用以衡量汉语二语学习者在写作中使用低频段（Bands 1-3，即常见和基础词汇）词汇的比例。汉语二语者的写作质量和写作分数越高，其作文中的低级词汇的比例则越低。换言之，LFPtr_L 越低，说明汉语二语者越能摆脱对基础词的依赖，更多运用复杂词汇来实现更精确和更丰富的表达。因此，LFPtr_L 的下降往往意味着写作者的语言水平和写作质量的提升。A 档作文的 LFPtr_L 平均值为 0.611，B 档作文的 LFPtr_L 平均值为 0.696，C 档作文的 LFPtr_L 平均值为 0.728。

h. Type ratio of LFP _ Middle-level (LFPtr_M) 和 Type ratio of LFP _ High-level (LFPtr_H) 与汉语二语写作质量呈显著正相关。汉语二语学习者写作质量和得分越高，就越容易在作文中使用更多中级和高级词汇，即复杂词汇。高分汉语二语作文的复杂词比例显著高于中低分组。复杂词比例是二语写作成熟度的重要表征之一。A 档作文的 LFPtr_M 平均值为 0.156, B 档作文的 LFPtr_M 平均值为 0.138, C 档作文的 LFPtr_M 平均值为 0.123; A 档作文的 LFPtr_H 平均值为 0.059，B 档作文的 LFPtr_H 平均值为 0.039，C 档作文的 LFPtr_H 平均值为 0.032。

i. Lexical Density (LD) 与汉语二语写作质量呈显著正相关。较高的词汇密度通常意味着文本信息量更大。二语者的写作质量和写作分数越高，LD 值越高，代表汉语二语作文的信息含量越丰富。在本研究中，我们将内容词划分为六类：除了名词、动词、形容词和副词外，还特别测量了汉语中常用的数词和量词的使用情况。作为孤立语，汉语在表达数量概念时，数词和名词之间必须通过量词连接，因此量词在汉语文本中占据重要地位。将数词和量词纳入词汇密度的考察，更能准确反映汉语写作的特点。因此高分的汉语二语作文往往展现出更高的词汇密度，这不仅说明其词汇知识更丰富，也一定程度上体现了他们对汉语特有结构（如量词系统）的更熟练掌握。A 档作文的 LD 平均值为 0.466，B 档作文的 LD 平均值为 0.441，C 档作文的 LD 平均值为 0.456。

j. CCC 与汉语二语写作质量呈显著正相关。该指标用以衡量学习者在写作中使用复杂汉字的比例。研究表明，汉语二语作文的质量和分数越高，往往伴随更高比例的复杂汉字使用。A 档作文的 CCC 平均值为 0.032，B 档作文的 CCC 平均值为 0.021，C 档作文的 CCC 平均值为 0.018。

k. Mean AoA Score (MAoAS) 与汉语二语的写作质量和写作分数呈显著正相关。水平更高的二语写作者在作文中往往使用习得年龄更晚的词汇。习得年龄更晚的词汇往往代表更复杂或抽象的概念，习得难度更高。随着汉语二语的写作质量和写作分数的提升，MAoAS 值变大，表明汉语二语写作者能够使用更多抽象、复杂的词汇，其词汇成熟度提高。A 档作文的 MAoAS 平均值为 9.644，B 档作文的 MAoAS 平均值为 9.423，C 档作文的 MAoAS 平均值为 9.289。

（2）在句法维度：主要考量句法复杂度。句法复杂度主要涵盖了语言产出的两个维度，即句法结构的复杂性和句法结构的多样性。对二语习得者来说，句法结构的复杂性和句法结构的多样性分别代表着二语句法知识的扩展范围和语言结构复杂变化的程度。此两者能够衡量二语者的语言行为及整体发展水平。句法复杂度的发展可以在某种程度上表明学习者能够使用更加多样化的句式和更为复杂及精细的语言项目。下面是具体的指标及其意义：

a. mean sentence length (MLS) 与汉语二语的写作质量和写作分数呈显著正相关。随着写作分数和质量提高，汉语二语者的 MSL 值不断增加，表明汉语二语者在语言产出中不仅具备更丰富的词汇资源，也能够作文中产出更长的句子，从而获得更高的写作质量评价。A 档作文的 MLS 平均值为 32.991，B 档作文的 MLS 平均值为 27.694，C 档作文的 MLS 平均值为 26.437。

b. Mean Length of Clauses (MLC) 与汉语二语的写作质量和写作分数呈显著负相关。随着写作分数和质量提高，汉语二语者的 MLC 值不断变短。汉语句式常采用散句、松句，汉语的分句或流水句较多，呈现出流散和疏放特征。换言之，高水平的汉语二语写作者不是单纯依赖长而复杂的从句结构，而是采用汉语的零句特征。A 档作文的 MLC 平均值为 11.116，B 档作文的 MLC 平均值为 11.394，C 档作文的 MLC 平均值为 11.706。

c. Number of Clauses Per Sentence (NCPS) 与汉语二语的写作质量和写作分数呈显著正相关。高水平汉语二语写作者的单个分句虽然较短，但往往能在一个句子中组织更多的分句，从而增强句子的层次性。这与汉语特有的“流水句”特征相呼应和“零句”特征相吻合。与 MLC 指标结合来看，表明 A 档作文更倾向于通过多个短分句的组合来完成语义的展开，而非依赖冗长的整句。这种写作方式更符合汉语意合的句法特征，即句子内部通过松散而灵活的分句并置来体现逻辑关系，从而实现自然流畅的表达。相比之下，低水平作文往往表现为分句较长、NCPS 较低，导致语义承载过于集中、句子显得冗长，影响表达的清晰性与自然度。A 档作文的 NCPS 平均值为 3.051，B 档作文的 NCPS 平均值为 2.480，C 档作文的 NCPS 平均值为 2.312。

d. 整体搭配多样性 (TOTAL_RTTR) 与汉语二语的写作质量和写作分数呈显著正相关。说明随着汉语二语写作分数和质量提高，汉语二语者在写作中倾向于使用更广泛的汉语句法搭配。A 档作文的 TOTAL_RTTR 平均值为 8.513，B 档作文的 TOTAL_RTTR 平均值为 7.815，C 档作文的 TOTAL_RTTR 平均值为 6.674。

e. 汉语特有搭配多样性 (UNIQUE_RTTR) 与汉语二语的写作质量和写作分数呈显著正相关。汉语特有搭配包括量名搭配、框式介词搭配、介动搭配和述补搭配。汉语特有句法搭配在二语写作评价中具有重要作用。汉语二语的写作质量和写作分数越高，汉语二语者更可能产出汉语特有搭配。A 档作文的 UNIQUE_RTTR 平均值为 2.542，B 档作文的 UNIQUE_RTTR 平均值为 2.099，C 档作文的 UNIQUE_RTTR 平均值为 1.712。

f. Verb-Object (VO_RTTR)、Subject-Predicate (SP_RTTR)、Adjective-Noun (AN_RTTR) 和 Adverb-Predicate (AP_RTTR) 四种跨语言通用搭配与 Classifier-Noun (CN_RTTR)、Preposition-Postposition (PP_RTTR)、Preposition-Verb (PV_RTTR) 和 Predicate-Complement (PC_RTTR) 四种汉语特有类型搭配均与汉语二语的写作质量和写作分数呈显著正相关。表明高水平的汉语二语者能在作文中产出更多样化的短语搭配。跨语言通用的搭配体现了写作者在二语习得中对基本语法关系的掌握，而汉语特有的搭配则反映了其对汉语结构特点的适应和熟练程度，语言特有特征对汉语二语学习者而言更为困难，因此需要更长的时间去掌握。在跨语言通用搭配上，A 档作文的 VO_RTTR 平均

值为 3.814, B 档作文的 VO_RTTR 平均值为 3.544, C 档作文的 VO_RTTR 平均值为 3.052; A 档作文的 SP_RTTR 平均值为 3.6370, B 档作文的 SP_RTTR 平均值为 3.468, C 档作文的 SP_RTTR 平均值为 3.039; A 档作文的 AN_RTTR 平均值为 1.743, B 档作文的 AN_RTTR 平均值为 1.493, C 档作文的 AN_RTTR 平均值为 1.217; A 档作文的 AP_RTTR 平均值为 5.453, B 档作文的 AP_RTTR 平均值为 5.029, C 档作文的 AP_RTTR 平均值为 4.216。在四种汉语特有类型搭配上, A 档作文的 CN_RTTR 平均值为 1.528, B 档作文的 CN_RTTR 平均值为 1.261, C 档作文的 CN_RTTR 平均值为 1.071; A 档作文的 PP_RTTR 平均值为 1.109, B 档作文的 PP_RTTR 平均值为 0.869, C 档作文的 PP_RTTR 平均值为 0.683; A 档作文的 PV_RTTR 平均值为 1.510, B 档作文的 PV_RTTR 平均值为 1.334, C 档作文的 PV_RTTR 平均值为 1.084; A 档作文的 PC_RTTR 平均值为 1.905, B 档作文的 PC_RTTR 平均值为 1.546, C 档作文的 PC_RTTR 平均值为 1.219。

g. VO_RATIO 和 SP_RATIO 与汉语二语的写作质量和写作分数呈显著负相关, 这意味着汉语二语作文成绩和作文质量越高, 动宾搭配和主谓搭配的使用比例越低。这是因为受汉语流水句特征的影响, 汉语主语和宾语的省略相对自由。高水平汉语二语写作者往往能够更自然地运用这种表达方式, 从而产出更符合汉语篇章风格的句子。A 档作文的 VO_RATIO 平均值为 0.209, B 档作文的 VO_RATIO 平均值为 0.218, C 档作文的 VO_RATIO 平均值为 0.223; A 档作文的 SP_RATIO 平均值为 0.183, B 档作文的 VO_RATIO 平均值为 0.196, C 档作文的 VO_RATIO 平均值为 0.207。

h. AN_RATIO 和 AP_RATIO 与汉语二语的写作质量和写作分数呈显著正相关, 这意味着汉语二语作文成绩和作文质量越高, 定中搭配和状中搭配的使用比例越高, 即修饰语使用越多, 意味着句子复杂度不断提升, 高水平汉语二语写作者往往能够灵活使用定语和状语来丰富句子内部的内嵌修饰关系。相比之下, 低水平写作者在修饰语的使用上相对受限, 句子更趋简单, 表达也显得单薄。A 档作文的 AN_RATIO 平均值为 0.045, B 档作文 AN_RATIO 平均值为 0.041, C 档作文 AN_RATIO 平均值为 0.040; A 档作文的 AP_RATIO 平均值为 0.399, B 档作文的 AP_RATIO 平均值为 0.402, C 档作文的 AP_RATIO 平均值为 0.387。

i. PP_RATIO、PV_RATIO 和 PC_RATIO 均与汉语二语的写作质量和写作分数呈显著正相关, 这意味着汉语二语写作成绩和质量越高, 汉语二语者就越能产出汉语特有的结构。A 档作文的 PP_RATIO 平均值为 0.027, B 档作文的 PP_RATIO 平均值为 0.021, C 档作文的 PP_RATIO 平均值为 0.020; A 档作文的 PV_RATIO 平均值为 0.038, B 档作文的 PV_RATIO 平均值为 0.037, C 档作文的 PV_RATIO 平均值为 0.036; A 档作文的 PC_RATIO 平均值为 0.054, B 档作文的 PC_RATIO 平均值为 0.045, C 档作文的 PC_RATIO 平均值为 0.042。

j. Mean dependency distance (MDD) 与汉语二语的写作质量和写作分数呈显著正相关。随着汉语二语写作分数和质量的提高, MDD 值不断升高。MDD 与人类的工作记忆和认知负荷密切相关, 是基于工作记忆容量限制的句法复杂度测量指标, 通常被认为可以用来衡量句法复杂度与语言加工的难度, 能够用来反映人类的语言结构和认知结构。这说明高质量和分数的汉语二语作文不再仅仅扩展句子的整体长度, 而是在作文中更有可能产出包含复杂句法结构的句子, 整体难度提升。A 档作文的 MDD 平均值为 2.969, B 档作文的 MDD 平均值为 2.736, C 档作文的 MDD 平均值为 2.684。

k. 话题链句比例 (NTCU_RATIO)、话题链平均依存距离 (TCMDD) 与话题链平均句长 (MLTTCU) 均与汉语二语的写作质量和写作分数呈显著正相关。汉语为话题突出性语言, 汉语允许主语省略, 以话题链作为测量汉语句法复杂度的指标, 在一定程度上能够反映汉语的语言类型学特征。也就是说, 随着汉语写作质量和分数的提高, 汉语二语者能够写出更多、更长、更复杂的汉语话题链, 越能产出具有汉语类型学特点的中介语。A 档作文的 NTCU_RATIO 平均值为 0.241, B 档作文的 NTCU_RATIO 平均值为 0.205, C 档作文的话 NTCU_RATIO 平均值为 0.184; A 档作文的 TCMDD 为 2.420, B 档作文的 TCMDD 为 2.257, C 档作文的 TCMDD 为 1.975; A 档作文的 MLTTCU 话题链平均句长为 14.799, B 档作文的 MLTTCU 话题链平均句长为 13.375, C 档作文的 MLTTCU 话题链平均句长为 11.528。

(3) 在语篇衔接维度: 衔接是语篇分析中的重要概念。衔接性对于展现二语语篇能力和写作质量的提升至关重要。我们主要关注能够体现汉语二语作文是否随着质量和分数的提升呈现汉语的类型学特征, 如主题突出和意合特征。下面是具体的指标及其意义:

a. 代词密度 (pron_density) 与汉语二语的写作质量和写作分数呈显著负相关。句均代词数 (pron_num_per_sent) 与汉语二语的写作质量和写作分数呈显著正相关。人称代词多样性 (per-pron_diversity) 与汉语二语的写作质量和写作分数呈显著正相关。这三个指标可以结合来看, 高水平汉语二语写作者往往减少冗余的代词使用, 但能在复杂句结构中更精准地使用代词, 并展现更大的多样性。这反映了他们能够在零照应和代词使用之间找到平衡。A 档作文的 pron_density 为 0.095, B 档作文的 pron_density 为 0.105, C 档作文的 pron_density 为 0.104; A 档作文的 pron_num_per_sent 为 1.876, B 档作文的 pron_num_per_sent 为 1.740, C 档作文的 pron_num_per_sent 为 1.779; A 档作文的 per-pron_diversity 为 0.177, B 档作文的 per-pron_diversity 为 0.178, C 档作文的 per-pron_diversity 为 0.155。

b. 主语密度 (subj_density) 与汉语二语的写作质量和写作分数呈显著负相关。代词主语多样性 (subj_pron_diversity) 与汉语二语的写作质量和写作分数呈显著负相关。这两个指标可以结合来分析, 写作质量和写作分数越高, subj_density 和 subj_pron_diversity 越低, 表明高水平汉语二语者更多使用隐含主语, 而不是显性地表达出来, 即通过零照应减少冗余从而提升汉语文本的衔接性, 越来越符合汉语话题突出的特点。A 档作文的 subj_density 为 0.099, B 档作文的 subj_density 为 0.107, C 档作文的 subj_density 为 0.109; A 档作文的 subj_pron_diversity 为 0.471, B 档作文的 subj_pron_diversity 为 0.449, C 档作文的 subj_pron_diversity 为 0.520。

c. 连词密度 (conj_density) 与汉语二语的写作质量和写作分数呈显著负相关。句均连词数 (conj_num_per_sent) 与汉语二语的写作质量和写作分数呈显著正相关。这两个指标可以结合来分析。汉语强调意合, 只要句子合乎逻辑, 汉语词汇、短语和从句之间不需要连词就可以按照顺序排列。因此, 高水平汉语二语写作者往往减少整体连词的密集使用, 但能在合适的句子结构中灵活而有效地运用连词, 从而实现符合汉语意合特征的自然连贯, 同时保持逻辑清晰和语篇层次感。A 档作文的 conj_density 为 0.042, B 档作文的连词密度平均值为 0.044, C 档作文的连词密度平均值为 0.044; A 档作文的 conj_num_per_sent 为 0.841, B 档作文的 conj_num_per_sent 为 0.771, C 档作文的

conj_num_per_sent 为 0.754。

(4) 在语言正确度维度：语言准确性是衡量汉语二语写作质量的重要维度。下面是具体的指标及其意义：

(标准化)字词句篇章偏误总数 (norm_total_errors) 、(标准化)字偏误数数量 (norm_chi_character_errors) 、(标准化)词偏误数数量 (norm_chi_word_errors) 、(标准化)句偏误数数量 (norm_sentence_errors) 、(标准化)篇章偏误数数量 (norm_discourse_errors) 均与汉语二语的写作质量和写作分数呈显著负相关。即偏误越多, 写作质量越低, 分数也越低。高质量的作文往往表现为偏误数量更少, 具有更高的语言准确度。A 档作文的 norm_total_errors 为 10.920, B 档作文的 norm_total_errors 为 10.979, C 档作文的 norm_total_errors 为 14.175; A 档作文的 norm_chi_character_errors 为 6.710, B 档作文的 norm_chi_character_errors 为 8.166, C 档作文的 norm_chi_character_errors 为 8.444; A 档作文的 norm_chi_word_errors 为 1.953, B 档作文的 norm_chi_word_errors 为 2.627, C 档作文的 norm_chi_word_errors 为 3.401; A 档作文的 norm_sentence_errors 为 0.803, B 档作文的 norm_sentence_errors 为 1.429, C 档作文的 norm_sentence_errors 为 2.184; A 档作文的 norm_discourse_errors 为 0.0578, B 档作文的 norm_discourse_errors 为 0.100, C 档作文的 norm_discourse_errors 为 0.145。

4.你需要严格遵守：

- (1) 评分时，必须评定为 high / medium / low 三个等级之一。
- (2) 评审反馈包括评分理由和修改意见，其中评分理由需要结合【作文内容】和【这篇作文对应的语言学指标数据】去解释你对这篇作文的评分，给出评分理由。修改意见主要是指出该作文不足之处及提升方向。
- (3) 给出评审反馈时，必须从 1.内容回应与交际功能、2.词汇、3.句法、4.语篇衔接、5.语言正确度这五点来输出评分理由以及对应的修改意见。尽可能融合更多有效的指标来分析，不能太简短。
- (4) 输出格式固定为 JSON，共 6 个键值对：{"1.评分": "high / medium / low", "2.评审反馈-内容回应与交际功能": "请结合内容回应与交际功能输出你对这篇作文的评审反馈", "3.评审反馈-词汇": "请结合词汇输出你对这篇作文的评审反馈", "4.评审反馈-句法": "请结合句法输出你对这篇作文的评审反馈", "5.评审反馈-语篇衔接": "请结合语篇衔接输出你对这篇作文的评审反馈", "6.评审反馈-语言正确度": "请语言正确度输出你对这篇作文的评审反馈"}

作文如下：

[题目]：TITLE

[作文内容]：CONTENT

示例

作文如下：

[题目]：一封求职信

[作文内容]：李先生：我从北京日报招聘启事中得悉 贵公司正招聘广告设计师，故特来信应聘。我在一九八九年毕业于香港理工学院，取得美术设计学院的高级文凭。在校内我是设计系系会的代表，经常参加广告公司协会举办的交流和讲座，使我对广告业的兴趣大大提高，也因此令我在广告设计方面奠下了良好的基础。毕业后我随即加入了李奥贝纳广告公司担任美术设计员，主要职责是正稿製作，製作范围包括报章杂志广告、还报、宣传单张及电视广告故事图。由于李奥贝纳是一间规模庞大的跨国公司，所以我接触到的客户面也比较广，计有国泰航空公司、雀巢饮品公司和香格里拉饭店等，而这些客户的服务和产品都跟旅游业有着密切的关系。在我四年的美术设计员工作中，我不但熟练了各类型广告宣传品的製作，更在美术效果的审视、广告材料的配合和意念的表达上获得不浅的经验。随函谨附简历表一份及近照一张，敬请察阅。如蒙安排面试，请电 2017531 转 2131 与我联络。联系地址：北京海淀区学院路 15 号公寓楼 303 室。敬祝业务蒸蒸日上
应聘者九三年十二月二十日

[这篇作文的语言学指标结果]：

- 一、词汇维度：NDW: 167; Uber: 93.515; VV: 0.3462; VV1: 0.8654; SVV1: 38.9423; CVV1: 4.4126; FWR_conjunction: 0.0568; Word Entropy: 6.9918; POS Entropy: 3.2291; MATTR-50: 0.8432; LFPtr_L: 0.4491; LFPtr_M: 0.2036; LFPtr_H: 0.0479; LD: 0.6201; CCC: 0.0588; Mean AoA Score (MAoAS): 11.2865
- 二、句法维度：MLS: 38.9; MLC: 16.913; NCPS: 2.3; TOTAL_RTTR: 7.2111; UNIQUE_RTTR: 2.4025; VO_RTTR: 3.873; SP_RTTR: 3.1623; AN_RTTR: 2.2361; AP_RTTR: 3.4641; CN_RTTR: 1.4142; PP_RTTR: 1.7321; PV_RTTR: 1.4142; PC_RTTR: 1.7321; VO_RATIO: 0.2885; SP_RATIO: 0.1923; AN_RATIO: 0.0962; AP_RATIO: 0.2308; PP_RATIO: 0.0577; PV_RATIO: 0.0385; PC_RATIO: 0.0577; MDD: 3.0568; NTCU_RATIO: 0.3; TCMD: 2.7097; MLTTCU: 21.6667

三、语篇衔接维度: pron_density: 0.0568; pron_num_per_sent: 1.3; per-pron_diversity: 0.0476; subj_density: 0.0699; subj_pron_diversity: 0.1667; conj_density: 0.0568; conj_num_per_sent: 1.3

四、语言正确度维度: norm_total_errors: 6.5502; norm_Chi_character_errors: 3.4934; norm_Chi_word_errors: 3.0568; norm_sentence_errors: 0.0; norm_discourse_errors: 0.0

Advanced Prompt with Scoring Rubric and Linguistic Features for Automated Writing Evaluation (in English)

You are now an expert in scoring HSK essays written by L2 Chinese learners. Please strictly follow the rules below to assign a score and provide evaluative comments:

- 1.Read the essay title and the full text carefully to identify the writing theme and direction.
- 2.Based on the performance of the essay, determine the proficiency level in accordance with the HSK writing scoring standards.

(1) Basic Level (C, 40–59 points): The essay content is insufficient, and the language is disorganized. Vocabulary is limited, with inappropriate word usage; grammar and character errors occur frequently, seriously affecting coherence and expression of ideas. Although the writer can express basic thoughts, overall communicative effectiveness is hindered.

(2) Intermediate Level (B, 60–79 points): The essay content is relatively complete, with a basically clear structure, and the writer can express ideas with some coherence. The writer demonstrates a certain command of grammar and vocabulary, with the ability to use some complex sentence patterns, though with limitations. Occasional inappropriate word usage, as well as grammar, vocabulary, and character errors, still occur, but overall communicative function is not impaired.

(3) Advanced Level (A, 80–100 points): The essay content is substantial, the structure is clear and logical, grammar is correct, and the language is fluent. The writer can skillfully use complex sentence structures and rhetorical devices, with a large vocabulary used appropriately and accurately. Chinese characters and punctuation are basically correct, with only a few minor errors that do not affect expression. Overall, the essay demonstrates strong logical organization and expressive power.
- 3.In addition, when assigning scores and providing evaluative comments, you must take into account both the essay content and the linguistic indice data corresponding to the essay, interpreted in light of the significance of each dimension of the indicators. As far as possible, the rationale for the assigned proficiency level should explicitly incorporate the results of these indicators, while also offering a detailed account of the weaknesses in each dimension and suggesting directions for improvement. The interpretations of the indices and the mean values for each proficiency band, presented below, are based on the aggregate results of 2,000 essays per level and may serve as reference benchmarks. It should be emphasized that these indicators do not necessarily exhibit monotonic linear variation (i.e., they may not consistently increase or decrease with proficiency level). Certain irregularities may arise from learners’ overuse or underuse of particular linguistic features. Therefore, in making evaluations, you must analyze the actual performance of the essay in conjunction with its indice data, avoiding mechanical reliance on the absolute value of any single indicator, and instead focusing on whether the overall pattern is consistent with the general level of writing proficiency.

(1)Lexical Dimension: Lexical richness refers to the quality of vocabulary knowledge demonstrated in a text and is an important dimension for assessing the language proficiency of second language learners. Lexical richness generally encompasses three sub-dimensions: lexical diversity, lexical sophistication, and lexical density. The specific indicators and their interpretations are presented below:

a. NDW shows a significant positive correlation with Chinese L2 writing scores and writing quality. A higher NDW value indicates that Chinese L2 learners are able to employ a greater variety of vocabulary in their writing. The average NDW value is 149.041 for Level A essays, 125.773 for Level B essays, and 96.715 for Level C essays.

b. Uber exhibits a significant positive correlation with Chinese L2 writing scores and writing quality. A higher Uber value suggests that learners are able to avoid excessive repetition of vocabulary and demonstrate a more mature ability to manage lexical resources, thereby achieving higher writing scores. The average Uber value is 59.639 for Level A essays, 52.806 for Level B essays, and 51.473 for Level C essays.

c. Verb-related indices include Verb Variation (VV), Verb Variation1 (VV1), Squared VV1 (SVV1), and Corrected VV1 (CVV1), all of which show significant positive correlations with Chinese L2 writing scores and writing quality. Verbs are the core components of a sentence, determining its basic structure and meaning. A higher level of verb variation indicates that learners are able to avoid repeated use of a limited set of verbs, indicating that learners are capable of making more fine-grained distinctions in their declarative sentences. Previous research has pointed out that, since Chinese is a verb-oriented language, verbs occupy a central position in the linguistic system and perform diverse functions. Therefore, these verb variation indices can effectively reflect the writing proficiency of Chinese L2 learners, as essays with high quality typically contain a greater diversity of verb usage. The average VV value is 0.403 for Level A essays, 0.396 for Level B essays, and 0.391 for Level C essays. The average VV1 value is 0.680 for Level A essays, 0.647 for Level B essays, and 0.670 for Level C essays. The average SVV1 value is 30.377 for Level A essays, 23.525 for Level B essays, and 17.956 for Level C essays. The average CVV1 value is 3.849 for Level A essays, 3.383 for Level B essays, and 2.948 for Level C essays.

d. FWR_conjunction shows a significant negative correlation with Chinese L2 writing scores and writing quality. As Chinese is a paratactic language, it primarily relies on implicit semantic connections rather than the extensive use of explicit

conjunctions to maintain logical relations. Excessive reliance on conjunctions suggests that Chinese learners depend heavily on hypotactic means, indicating that their Chinese expression is not yet fully mature. The average FWR_conjunction value is 0.042 for Level A essays, 0.044 for Level B essays, and 0.044 for Level C essays.

e. Word Entropy and POS Entropy both show significant positive correlations with Chinese L2 writing scores and writing quality. Word Entropy indices the uniformity and uncertainty of information distribution in a text. A higher entropy value indicates a more balanced distribution of vocabulary, less excessive repetition, and greater flexibility in language organization. In Chinese L2 writing, a higher Word Entropy value typically suggests that learners are able to draw on a wider range of lexical resources. Similarly, a higher POS Entropy value indicates that learners are able to employ a variety of word classes (e.g., nouns, verbs, adjectives, adverbs, numerals, and classifiers), leading to greater syntactic flexibility. The average Word Entropy value is 6.698 for Level A essays, 6.452 for Level B essays, and 6.118 for Level C essays. The average POS Entropy value is 3.155 for Level A essays, 3.128 for Level B essays, and 3.103 for Level C essays.

f. MATTR-50 shows a significant positive correlation with Chinese L2 writing scores and writing quality. MATTR-50 is used to indice the overall lexical diversity of a text and provides a relatively stable assessment of this feature. In Chinese L2 writing, a higher MATTR-50 value generally indicates that learners employ a richer and more diverse vocabulary in their writing, which in turn leads to higher writing scores and quality evaluations. The average MATTR-50 value is 0.799 for Level A essays, 0.776 for Level B essays, and 0.758 for Level C essays.

g. Type ratio of LFP – Low-level (LFPtr_L) shows a significant negative correlation with Chinese L2 writing quality. It evaluates the proportion of low-frequency-band vocabulary (Bands 1–3, i.e., common and basic words) used in writing. Higher writing quality and scores are associated with a lower proportion of low-level vocabulary. A lower LFPtr_L value therefore indicates less reliance on basic vocabulary and greater use of complex words for more precise and varied expression. A decrease in LFPtr_L often reflects improvement in writing proficiency. The average LFPtr_L value is 0.611 for Level A essays, 0.696 for Level B essays, and 0.728 for Level C essays.

h. LFPtr_M and LFPtr_H show significant positive correlations with Chinese L2 writing quality. Chinese L2 Learners with higher writing quality and scores tend to use a larger proportion of mid-level and high-level vocabulary, i.e., more complex words. High-scoring essays contain a significantly greater proportion of complex vocabulary than mid-scoring and low-scoring ones. The proportion of complex words is considered an important indicator of maturity in second language writing. The average LFPtr_M value is 0.156 for Level A essays, 0.138 for Level B essays, and 0.123 for Level C essays. The average LFPtr_H value is 0.059 for Level A essays, 0.039 for Level B essays, and 0.032 for Level C essays.

i. LD shows a significant positive correlation with Chinese L2 writing quality. A higher lexical density usually indicates that a text conveys more information. The higher the writing quality and scores of Chinese L2 learners, the higher the LD value, reflecting a richer informational content in their essays. In this study, content words were classified into six categories. In addition to nouns, verbs, adjectives, and adverbs, we also specifically indiced the use of numerals and classifiers, which are particularly salient in Chinese. As an isolating language, Chinese requires classifiers to connect numerals and nouns when expressing quantity concepts, making classifiers an essential component in Chinese texts. Including numerals and classifiers in the indicement of lexical density provides a more accurate reflection of the characteristics of Chinese writing. Accordingly, high-scoring Chinese L2 essays typically exhibit higher lexical density, which not only demonstrates richer lexical knowledge but also reflects, to some extent, learners' greater mastery of language-specific structures such as the classifier system. The average LD value is 0.466 for Level A essays, 0.441 for Level B essays, and 0.456 for Level C essays.

j. CCC shows a significant positive correlation with Chinese L2 writing quality. It evaluates the proportion of complex Chinese characters used in learner writing. Research indicates that Chinese L2 essays with higher quality and score are typically associated with a greater proportion of complex characters. The average CCC value is 0.032 for Level A essays, 0.021 for Level B essays, and 0.018 for Level C essays.

k. MAoAS shows a significant positive correlation with Chinese L2 writing quality and writing scores. More advanced L2 writers tend to use words acquired at a later age. Such words often represent more complex or abstract concepts and are generally more difficult to learn. As Chinese L2 writing quality and scores increase, the MAoAS value also rises, indicating that learners are able to employ a greater range of abstract and complex vocabulary, reflecting higher lexical maturity. The average MAoAS value is 9.644 for Level A essays, 9.423 for Level B essays, and 9.289 for Level C essays.

(2) Syntactic Dimension: This dimension primarily examines syntactic complexity, which encompasses the sophistication of syntactic structures and the diversity of syntactic structures. For L2 learners, syntactic sophistication and syntactic diversity respectively represent the breadth of their syntactic knowledge and the degree of structural complexity and variation in their language use. Together, these aspects provide an important indice of learners' linguistic performance and overall developmental level. The development of syntactic complexity can, to some extent, indicate learners' ability to employ a wider variety of sentence patterns as well as more complex and refined linguistic forms. The specific indices and their interpretations are presented below:

a. MSL shows a significant positive correlation with Chinese L2 writing quality and writing scores. As writing quality and scores increase, the MSL value of Chinese L2 learners also rises, indicating that they not only possess richer lexical resources in language production but are also able to produce longer sentences in their essays, thereby achieving higher evaluations of writing quality. The average MSL value is 32.991 for Level A essays, 27.694 for Level B essays, and 26.437 for Level C essays.

b. MLC shows a significant negative correlation with Chinese L2 writing quality and writing scores. As writing quality and scores increase, the MLC value of Chinese L2 learners decreases. Chinese sentences often employ parataxis and looser

sentence structures, with a greater prevalence of loosely connected or sequential clauses, reflecting a tendency toward dispersiveness and openness in structure. In other words, more proficient Chinese L2 writers do not simply rely on long and complex subordinate clause structures but instead adopt features such as the use of elliptical sentences, which are characteristic of Chinese. The average MLC value is 11.116 for Level A essays, 11.394 for Level B essays, and 11.706 for Level C essays.

c. NCPS shows a significant positive correlation with Chinese L2 writing quality and writing scores. Although the individual clauses produced by more advanced Chinese L2 writers tend to be shorter, they are often able to incorporate a greater number of clauses within a single sentence, thereby enhancing its hierarchical structure. This corresponds to the characteristic of “run-on sentences” (liúshuǐjù) in Chinese and aligns with the use of “elliptical sentences” (língjù). In conjunction with the MLC, this indicates that Level A essays are more likely to develop meaning through the combination of multiple short clauses rather than relying on lengthy single sentences. Such a writing strategy better reflects the paratactic syntactic features of Chinese, in which logical relations are expressed through the loose and flexible juxtaposition of clauses, resulting in natural and fluent expression. By contrast, essays at lower proficiency levels tend to have longer clauses and lower NCPS values, leading to an overly concentrated semantic load and making sentences appear lengthy, which reduces clarity and naturalness of expression. The average NCPS value is 3.051 for Level A essays, 2.480 for Level B essays, and 2.312 for Level C essays.

d. TOTAL_RTTR shows a significant positive correlation with Chinese L2 writing quality and writing scores. This suggests that as the quality and scores of Chinese L2 writing improve, learners are more likely to make use of a broader range of syntactic combinations in their compositions. The average TOTAL_RTTR value is 8.513 for Level A essays, 7.815 for Level B essays, and 6.674 for Level C essays.

e. UNIQUE_RTTR shows a significant positive correlation with Chinese L2 writing quality and writing scores. Chinese specific combinations include Classifier-Noun combinations, Preposition-Postposition combinations, preposition-verb combinations, and Predicate-Complement combinations. Such combinations play an important role in the evaluation of Chinese L2 writing. The higher the quality and scores of Chinese L2 writing, the more likely learners are to produce Chinese-specific combinations. The average UNIQUE_RTTR value is 2.542 for Level A essays, 2.099 for Level B essays, and 1.712 for Level C essays.

f. Cross-linguistically common combinations, including Verb-Object (VO_RTTR), Subject-Predicate (SP_RTTR), Adjective-Noun (AN_RTTR), and Adverb-Predicate (AP_RTTR), as well as Chinese specific combinations, including Classifier-Noun (CN_RTTR), Preposition-Postposition (PP_RTTR), Preposition-Verb (PV_RTTR), and Predicate-Complement (PC_RTTR), all show significant positive correlations with Chinese L2 writing quality and writing scores. This indicates that more proficient Chinese L2 writers are able to produce a greater variety of phrasal combinations in their essays. Cross-linguistically common combinations reflect learners’ mastery of basic grammatical relations in second language acquisition, whereas Chinese-specific combinations capture their adaptation to and proficiency with the structural characteristics of Chinese. Since language-specific features are generally more difficult for Chinese L2 learners, they require a longer period of acquisition. For cross-linguistically common combinations, the average VO_RTTR value is 3.814 for Level A essays, 3.544 for Level B essays, and 3.052 for Level C essays; the average SP_RTTR value is 3.637 for Level A essays, 3.468 for Level B essays, and 3.039 for Level C essays; the average AN_RTTR value is 1.743 for Level A essays, 1.493 for Level B essays, and 1.217 for Level C essays; and the average AP_RTTR value is 5.453 for Level A essays, 5.029 for Level B essays, and 4.216 for Level C essays. For Chinese-specific combinations, the average CN_RTTR value is 1.528 for Level A essays, 1.261 for Level B essays, and 1.071 for Level C essays; the average PP_RTTR value is 1.109 for Level A essays, 0.869 for Level B essays, and 0.683 for Level C essays; the average PV_RTTR value is 1.510 for Level A essays, 1.334 for Level B essays, and 1.084 for Level C essays; and the average PC_RTTR value is 1.905 for Level A essays, 1.546 for Level B essays, and 1.219 for Level C essays.

g. VO_RATIO and SP_RATIO show significant negative correlations with Chinese L2 writing quality and writing scores. This means that the higher the writing scores and quality of Chinese L2 essays, the lower the proportion of Verb-Object and Subject-Predicate collocations. This pattern can be attributed to the influence of the “run-on sentence” characteristic in Chinese, where the omission of subjects and objects is relatively flexible. More proficient Chinese L2 writers are often able to employ this feature more naturally, thereby producing sentences that better conform to the discourse style of Chinese. The average VO_RATIO value is 0.209 for Level A essays, 0.218 for Level B essays, and 0.223 for Level C essays; the average SP_RATIO value is 0.183 for Level A essays, 0.196 for Level B essays, and 0.207 for Level C essays.

h. AN_RATIO and AP_RATIO show significant positive correlations with Chinese L2 writing quality and writing scores. This means that the higher the scores and quality of Chinese L2 essays, the higher the proportion of Adjective-Noun and Adverb-Predicate combinations. In other words, the increased use of modifiers indicates greater syntactic complexity. More proficient Chinese L2 writers are often able to employ attributive and adverbial modifiers flexibly to enrich the embedded modification relationships within sentences. By contrast, less proficient writers tend to use modifiers in a more limited way, resulting in simpler sentences and thinner expression. The average AN_RATIO value is 0.045 for Level A essays, 0.041 for Level B essays, and 0.040 for Level C essays; the average AP_RATIO value is 0.399 for Level A essays, 0.402 for Level B essays, and 0.387 for Level C essays.

i. PP_RATIO, PV_RATIO, and PC_RATIO all show significant positive correlations with Chinese L2 writing quality and writing scores. This means that the higher the scores and quality of Chinese L2 essays, the more likely learners are to produce Chinese-specific structures. The average PP_RATIO value is 0.027 for Level A essays, 0.021 for Level B essays, and 0.020 for Level C essays; the average PV_RATIO value is 0.038 for Level A essays, 0.037 for Level B essays, and 0.036 for

Level C essays; and the average PC_RATIO value is 0.054 for Level A essays, 0.045 for Level B essays, and 0.042 for Level C essays.

j. MDD shows a significant positive correlation with Chinese L2 writing quality and writing scores. As writing scores and quality increase, the MDD value also rises. MDD is closely related to human working memory and cognitive load, serving as a syntactic complexity indice grounded in the capacity limitations of working memory. It is generally regarded as a indice for assessing syntactic complexity and the difficulty of language processing, and it reflects both linguistic and cognitive structures. This suggests that higher-quality and higher-scoring Chinese L2 essays do not merely extend overall sentence length but are more likely to produce sentences containing complex syntactic structures, thereby increasing overall difficulty. The average MDD value is 2.969 for Level A essays, 2.736 for Level B essays, and 2.684 for Level C essays.

k. NTCU_RATIO, TCMDD, and MLTTCU all show significant positive correlations with Chinese L2 writing quality and writing scores. As a topic-prominent language, Chinese allows subject omission, and using topic chains as a indice of syntactic complexity provides an indication of the typological characteristics of Chinese. In other words, as the quality and scores of Chinese L2 writing improve, learners are able to produce more, longer, and more complex topic chains, thereby generating interlanguage that more closely reflects the Chinese typological features. The average NTCU_RATIO value is 0.241 for Level A essays, 0.205 for Level B essays, and 0.184 for Level C essays; the average TCMDD value is 2.420 for Level A essays, 2.257 for Level B essays, and 1.975 for Level C essays; and the average MLTTCU value is 14.799 for Level A essays, 13.375 for Level B essays, and 11.528 for Level C essays.

(3) Discourse Cohesion Dimension: Cohesion is an important concept in discourse analysis. Cohesiveness is crucial for demonstrating second language discourse competence and for improving writing quality. In this study, we focus primarily on indicators that capture whether Chinese L2 essays exhibit typological features of Chinese, such as topic prominence and parataxis, as their quality and scores increase. The specific indices and their interpretations are presented below:

(3) Discourse Cohesion Dimension: Cohesion is an important concept in discourse analysis. Cohesiveness is crucial for demonstrating second language discourse competence and for improving writing quality. In this study, we focus primarily on indices that capture whether Chinese L2 essays exhibit typological features of Chinese, such as topic prominence and parataxis, as their quality and scores increase. The specific indices and their interpretations are presented below:

a. pron_density shows a significant negative correlation with Chinese L2 writing quality and writing scores, while pron_num_per_sent and per-pron_diversity both show significant positive correlations. Taken together, these three indices suggest that more proficient Chinese L2 writers reduce redundant pronoun use while employing pronouns more precisely in complex sentence structures and with greater diversity. This reflects their ability to strike a balance between zero anaphora and pronoun use. The average pron_density value is 0.095 for Level A essays, 0.105 for Level B essays, and 0.104 for Level C essays; the average pron_num_per_sent value is 1.876 for Level A essays, 1.740 for Level B essays, and 1.779 for Level C essays; and the average per-pron_diversity value is 0.177 for Level A essays, 0.178 for Level B essays, and 0.155 for Level C essays.

b. subj_density shows a significant negative correlation with Chinese L2 writing quality and writing scores, and subj_pron_diversity also shows a significant negative correlation. Taken together, these two indices indicate that as writing quality and scores improve, both subj_density and subj_pron_diversity decrease. This suggests that more proficient Chinese L2 writers make greater use of implicit subjects rather than expressing them explicitly, thereby reducing redundancy through zero anaphora and enhancing cohesion in Chinese texts. Such usage increasingly aligns with the topic-prominent characteristics of Chinese. The average subj_density value is 0.099 for Level A essays, 0.107 for Level B essays, and 0.109 for Level C essays; the average subj_pron_diversity value is 0.471 for Level A essays, 0.449 for Level B essays, and 0.520 for Level C essays.

c. conj_density shows a significant negative correlation with Chinese L2 writing quality and writing scores, whereas conj_num_per_sent shows a significant positive correlation. Taken together, these two indices indicate that, because Chinese emphasizes parataxis, words, phrases, and clauses can be arranged sequentially without conjunctions as long as the sentence remains logically coherent. Accordingly, more proficient Chinese L2 writers tend to reduce the overall density of conjunction use but are able to employ conjunctions flexibly and effectively within appropriate sentence structures. This allows them to achieve natural cohesion consistent with the paratactic features of Chinese while maintaining logical clarity and discourse organization. The average conj_density value is 0.042 for Level A essays, 0.044 for Level B essays, and 0.044 for Level C essays; the average conj_num_per_sent value is 0.841 for Level A essays, 0.771 for Level B essays, and 0.754 for Level C essays.

(4) Linguistic Accuracy Dimension: Linguistic accuracy is an essential dimension for evaluating the quality of Chinese L2 writing. The specific indices and their interpretations are presented below:

norm_total_errors, norm_Chi_character_errors, norm_Chi_word_errors, norm_sentence_errors, and norm_discourse_errors all show significant negative correlations with Chinese L2 writing quality and writing scores. In other words, the greater the number of errors, the lower the writing quality and scores. High-quality essays are typically characterized by fewer errors and therefore higher levels of linguistic accuracy. The average norm_total_errors value is 10.920 for Level A essays, 10.979 for Level B essays, and 14.175 for Level C essays; the average norm_Chi_character_errors value is 6.710 for Level A essays, 8.166 for Level B essays, and 8.444 for Level C essays; the average norm_Chi_word_errors value is 1.953 for Level A essays, 2.627 for Level B essays, and 3.401 for Level C essays; the average norm_sentence_errors value is 0.803 for Level A essays, 1.429 for Level B essays, and 2.184 for Level C essays; and the average norm_discourse_errors value is 0.0578 for Level A essays, 0.100 for Level B essays, and 0.145 for Level C essays.

4. You must strictly follow these rules:

- (1) When assigning a score, you must rate the essay as one of three levels: high / medium / low.
- (2) The review must include both a scoring rationale and revision suggestions. The scoring rationale should integrate the essay content and the linguistic indicator data corresponding to the essay to justify your rating. The revision suggestions should primarily identify the weaknesses of the essay and provide directions for improvement.
- (3) When giving the review, you must provide both the scoring rationale and the corresponding revision suggestions from the following five aspects: 1. Content response and communicative function, 2. Vocabulary, 3. Syntax, 4. Discourse cohesion, 5. Linguistic accuracy. Try to incorporate as many relevant indicators as possible in your analysis, and avoid being overly brief.
- (4) The output format must be fixed in JSON with six key-value pairs:
- ```
{ "1. Score": "high / medium / low", "2. Feedback - Content Response and Communicative Function": "Please provide your feedback of the essay from the perspective of content response and communicative function", "3. Feedback - Vocabulary": "Please provide your feedback of the essay from the perspective of vocabulary", "4. Feedback - Syntax": "Please provide your feedback of the essay from the perspective of syntax", "5. Feedback - Discourse Cohesion": "Please provide your feedback of the essay from the perspective of discourse cohesion", "6. Feedback - Linguistic Accuracy": "Please provide your feedback of the essay from the perspective of linguistic accuracy"}
```

Essay:

[题目]: TITLE

[作文内容]: CONTENT

[The linguistic indice results of this essay]:

[METRICS]

for example

## Essay:

[title]: 一封求职信

[作文内容]: 李先生：我从北京日报招聘启事中得悉 贵公司正招聘广告设计师，故特来信应聘。我在一九八九年毕业于香港理工学院，取得美术设计学院的高级文凭。在校内我是设计系系的代表，经常参加广告公司协会举办的交流和讲座，使我对广告业的兴趣大大提高，也因此令我在广告设计方面奠下了良好的基础。毕业后我随即加入了李奥贝纳广告公司担任美术设计员，主要职责是正稿製作，製作范围包括报章杂志广告、还报、宣传单张及电视广告故事图。由于李奥贝纳是一间规模庞大的跨国公司，所以我接触到的客户面也比较广，计有国泰航空公司、雀巢饮品公司和香格里拉饭店等，而这些客户的服务和产品都跟旅游业有着密切的关系。在我四年的美术设计员工作中，我不但熟习了各类型广告宣传品的製作，更在美术效果的审视、广告材料的配合和意念的表达上获得不浅的经验。随函谨附简历表一份及近照一张，敬请察阅。如蒙安排面试，请电 2017531 转 2131 与我联络。联系地址：北京海淀区学院路 15 号公寓楼 303 室。敬祝业务蒸蒸日上应聘者九三年十二月二十日

[The linguistic indice results of this essay]:

1. Lexical Dimension: NDW: 167; Uber: 93.515; VV: 0.3462; VV1: 0.8654; SVV1: 38.9423; CVV1: 4.4126; FWR\_conjunction: 0.0568; Word Entropy: 6.9918; POS Entropy: 3.2291; MATTR-50: 0.8432; LFPtr\_L: 0.4491; LFPtr\_M: 0.2036; LFPtr\_H: 0.0479; LD: 0.6201; CCC: 0.0588; Mean AoA Score (MAoAS): 11.2865

2. Syntactic Dimension: MLS: 38.9; MLC: 16.913; NCPS: 2.3; TOTAL\_RTTR: 7.2111; UNIQUE\_RTTR: 2.4025; VO\_RTTR: 3.873; SP\_RTTR: 3.1623; AN\_RTTR: 2.2361; AP\_RTTR: 3.4641; CN\_RTTR: 1.4142; PP\_RTTR: 1.7321; PV\_RTTR: 1.4142; PC\_RTTR: 1.7321; VO\_RATIO: 0.2885; SP\_RATIO: 0.1923; AN\_RATIO: 0.0962; AP\_RATIO: 0.2308; PP\_RATIO: 0.0577; PV\_RATIO: 0.0385; PC\_RATIO: 0.0577; MDD: 3.0568; NTCU\_RATIO: 0.3; TCMDD: 2.7097; MLTTCU: 21.6667

3. Discourse Cohesion Dimension: pron\_density: 0.0568; pron\_num\_per\_sent: 1.3; per-pron\_diversity: 0.0476; subj\_density: 0.0699; subj\_pron\_diversity: 0.1667; conj\_density: 0.0568; conj\_num\_per\_sent: 1.3

4. Linguistic Accuracy Dimension: norm\_total\_errors: 6.5502; norm\_Chi\_character\_errors: 3.4934; norm\_Chi\_word\_errors: 3.0568; norm\_sentence\_errors: 0.0; norm\_discourse\_errors: 0.0

Appendix F. HSK meta rater prompt



Chinese version

你现在是汉语二语 HSK 作文评审主席，请严格按照以下规则对作文进行评分并给出评审反馈。必须参考【两位评分员的评分及评审反馈】，先总结和分析两位评分员的评分及评审反馈，再结合你自己的判断做最终决定：

(1-3 部分同 **Advanced Prompt** )

4. 你的任务：

(1) 必须参考【两位评分员的评分及评审反馈】，先总结和分析两位评分员的评分及评审意见，再结合你自己的判断做最终决定。

(2) 评分时，必须参考【两位评分员的评分及评审反馈】，再综合评定为 high/medium/low 三个等级之一。

(3) 评审反馈包括评分理由和修改意见。其中评分理由主要参考【两位评分员的评分及评审反馈】并结合【作文内容】和【这篇作文对应的语言学指标数据】综合地去解释你对这篇作文的评分。修改意见主要是指出该作文不足之处及提升方向。

(4) 给出评审反馈时，必须从 1.内容回应与交际功能、2.词汇、3.句法、4.语篇衔接、5.语言正确度这五点来输出评分理由以及对应的修改意见。必须参考【两位评分员的评分及评审反馈】进行总结和分析，再结合你自己的判断要给出综合性反馈，不能简短概述。

(5) 输出格式固定为 JSON，共 6 个键值对：{"1.评分": "high / medium / low", "2.评审反馈-内容回应与交际功能": "请输出你对这篇作文在内容回应与交际功能的综合评审反馈", "3.评审反馈-词汇": "请输出你对这篇作文在词汇的综合评审反馈", "4.评审反馈-句法": "请输出你对这篇作文在句法的综合评审反馈", "5.评审反馈-语篇衔接": "请输出你对这篇作文在语篇衔接的综合评审反馈", "6.评审反馈-语言正确度": "请输出你对这篇作文在语言正确度的综合评审反馈"}

## 作文如下：

[题目]: TITLE

[作文已知等级]: LEVEL

[作文内容]: CONTENT

[这篇作文的语言学指标结果]:

[METRICS]

【两位评分员的评分及评审意见】：

REVIEWERS\_PROMPT

**\*\*请注意，再次提示：评分和给出各个维度的评审意见时，必须参考【两位评分员的评分及评审意见】，先总结和分析两位评分员的评分及评审意见，再结合你自己的判断做最终决定。**

\*\*\*\*\*

English version

You are now the meta reviewer of the HSK essay review committee. Please strictly follow the rules below to assign a score and provide evaluation feedback on the essay. You must refer to [the two raters’ scores and evaluation feedback], first summarizing and analyzing their evaluations, and then making your own final decision:

(Parts 1-3 are the same as in the Advanced Prompt)

4. Your task:

(1) You must refer to [the two raters’ scores and review feedback], first summarizing and analyzing their evaluations, and then making your own final decision.

(2) When assigning the score, you must take into account [the two raters’ scores and evaluation feedback] and then give a comprehensive rating of one of three levels: high/medium/low.

(3) The review feedback must include both a scoring rationale and revision suggestions. The scoring rationale should primarily be based on [the two raters’ scores and review feedback] and then integrate both the essay content and the linguistic indicator data of the essay to explain your scoring decision. The revision suggestions should focus on identifying the weaknesses of the essay and providing directions for improvement.

(4) When giving the evaluation feedback, you must address all five dimensions: 1. Content Response and Communicative Function, 2. Vocabulary, 3. Syntax, 4. Discourse Cohesion, 5. Linguistic Accuracy. You must summarize and analyze [the

two raters’ scores and evaluation feedback], then combine this with your own judgment to provide a comprehensive review. Simple summaries are not acceptable.

(5) The output format must be fixed in JSON with six key-value pairs:

```
{ "1. Score": "high / medium / low", "2. Review - Content Response and Communicative Function": "Please provide your comprehensive review of the essay from the perspective of content response and communicative function", "3. Review - Vocabulary": "Please provide your comprehensive review of the essay from the perspective of vocabulary", "4. Review - Syntax": "Please provide your comprehensive review of the essay from the perspective of syntax", "5. Review - Discourse Cohesion": "Please provide your comprehensive review of the essay from the perspective of discourse cohesion", "6. Review - Linguistic Accuracy": "Please provide your comprehensive review of the essay from the perspective of linguistic accuracy"}
```

```
##Essay:
[题目]: TITLE
[作文已知等级]: LEVEL
[作文内容]: CONTENT
[The linguistic indice results of this essay]:
[METRICS]
[The two raters’ scores and review feedback]:
REVIEWERS_PROMPT

Please note, once again: When assigning the score and providing evaluation feedback for each dimension, you must refer to [the two raters’ scores and evaluation feedback], first summarizing and analyzing their evaluations, and then making your own final decision.
```

Appendix G The prompt for instructing Qwen3-80B to generate synthetic evaluation feedback annotations

Chinese version

```

（1-3 部分同 Advanced Prompt ）

4. 你的任务：

（1）你不需要评分，而是根据作文的已知等级（LEVEL 等级），结合其原文以及对应维度的指标数据，给出各维度的评审反馈。评审反馈包括评分理由和修改意见。

（2）给出评分理由时，你需要假装不知道这篇作文的等级（实际上你知道），你需要先基于这篇作文的内容和指标数据进行分析，然后再给出与作文等级之间关系的结论。简单来说，单纯看你的评分理由需要让人觉得你不知道这篇作文的等级，你現在就在推测该作文等级。尽可能融合更多有效的指标来分析，不能太简短。

（3）给出修改意见时，你需要基于作文内容、指标数据以及对应维度的评分理由，详细地指出不足之处并给出针对性意见和提升的方向。尽可能融合更多有效的指标来分析，不能太简短。

（4）给出评审反馈时，必须从： 1.内容回应与交际功能、2.词汇、3.句法、4.语篇衔接、5.语言正确度这五点来输出评分理由以及对应的修改意见。尽可能融合更多有效的指标特征来分析，不能太简短。

（5）输出格式固定为 json，共 6 个键值对：{"1.作文已知等级": "该作文已知等级", "2.评审反馈-内容回应与交际功能": "请结合内容回应与交际功能输出你对这篇作文的评审反馈", "3.评审反馈-词汇": "请结合词汇指标特征输出你对这篇作文的评审反馈", "4.评审反馈-句法": "请结合句法指标特征输出你对这篇作文的评审反馈", "5.评审反馈-语篇衔接": "请结合语篇衔接指标特征输出你对这篇作文的评审反馈", "6.评审反馈-语言正确度": "请结合语言正确度指标特征输出你对这篇作文的评审反馈"}
```

```

作文如下：
```

[题目]: TITLE

[作文已知等级]: LEVEL

[作文内容]: CONTENT

[这篇作文的语言学指标结果]:

[METRICS]

English version

(Parts 1-3 are the same as the Advanced Prompt)

4. Your task:

- (1) You do not need to assign a score. Instead, based on the known level of the essay (LEVEL), together with the essay content and its corresponding indice data, you should provide review comments for each dimension. The review comments must include both the scoring rationale and revision suggestions.
- (2) When providing the scoring rationale, you should pretend not to know the assigned level of the essay (even though you actually do). You must first analyze the essay content and the indice data, and only then draw a conclusion about how these relate to the essay’s level. Put simply, your rationale should read as if you were unaware of the essay’s level and are currently inferring it. Incorporate as many relevant indices as possible in your analysis, and avoid being overly brief.
- (3) When providing revision suggestions, you should base them on the essay content, the indice data, and your rationale for the corresponding dimension. You need to identify weaknesses in detail and give targeted suggestions for improvement, again incorporating as many relevant indices as possible, and avoid being overly brief.
- (4) When giving review comments, you must address all five dimensions: 1. Content Response and Communicative Function, 2. Vocabulary, 3. Syntax, 4. Discourse Cohesion, 5. Linguistic Accuracy. For each dimension, provide both the scoring rationale and corresponding revision suggestions. Incorporate as many relevant linguistic indices as possible in your analysis, and avoid being overly brief.
- (5) The output format must be fixed in JSON with six key-value pairs: { "1. Known Level of the Essay": "The given level of the essay", "2. Feedback - Content Response and Communicative Function": "Please provide your feedback of the essay from the perspective of content response and communicative function", "3. Feedback - Vocabulary": "Please provide your feedback of the essay from the perspective of vocabulary indices", "4. Feedback - Syntax": "Please provide your feedback of the essay from the perspective of syntactic indices", "5. Feedback - Discourse Cohesion": "Please provide your feedback of the essay from the perspective of discourse cohesion indices", "6. Feedback - Linguistic Accuracy": "Please provide your feedback of the essay from the perspective of linguistic accuracy indices"}

##Essay:

[题目]: TITLE

[作文已知等级]: LEVEL

[作文内容]: CONTENT

[The linguistic indice results of this essay]:

[METRICS]

Appendix H. Prompt for examining Model-Generated Feedback Performance

Chinese version

请你现在作为 AWE 反馈的评估者。你的任务是根据以下标准，评估某位评分员的评分与评审反馈是否合理。具体来说，你需要结合作文的原始人工等级、语言学指标结果和客观评分标准，从以下三个角度进行判断：

角度一【客观证据覆盖度】：评估各维度的评审反馈是否用到了可复核的证据，以及统计表述是否正确。

- 5分 | 完备可复核：**明确输出相关语言学指标并给出具体数值，比较词的方向正确，如使用“高/低/远高/远低”，能结合评分标准、提供基线；需要文本证据的给出可定位的语句或片段；结论与证据一一对应，可被复算重现。
- 4分 | 证据充分、少量缺项：**指标与数值齐全、方向正确；但是要么缺评分标准/基线、要么缺文本定位，不过不影响主要论断的复核。
- 3分 | 部分证据：**提及指标但缺数值或只给相对词（较高/较低）；或给出数值但无基线却使用了比较性措辞；方向基本正确且无重大统计错误。
- 2分 | 证据薄弱/统计可疑：**只有笼统判断或仅有单一指标支撑复杂结论；出现夸大用语但无任何阈值；或出现一个重要方向性错误（高低写反等）。
- 1分 | 几乎无证据：**纯主观判断，无根据；结论与数据显著不符或虚构基线；即便出现数字也来源不明/口径不清或多处统计表述错误，难以复核。

**角度二【内部一致性】：**评估各维度评审反馈以及综合维度评审反馈中，同一段或同一条评语内部的逻辑连贯性、术语使用一致性，以及证据与结论之间的一致性。

- 5分 | 完全一致：**论断之间相互支持，无矛盾或张冠李戴。
- 4分 | 基本一致：**偶尔出现轻微夸张或表述模糊，但核心逻辑未受影响；术语和范围总体一致。
- 3分 | 局部矛盾，可修正：**出现少量可定位的不一致（如一会说“连词极少”，一会又说“完全没有”），但核心结论仍成立。
- 2分 | 多处不一致/跳步推断：**多处矛盾或关键逻辑跳跃而无过渡证据。
- 1分 | 严重自相矛盾：**逻辑无法修补，结论之间彼此否定；或段内前后使用相反论断导致整体不可解释。

**角度三【整体修改意见质量】：**评估各维度评审反馈语气是否积极、尊重且建设性强，既包含真诚的正向肯定，又能提出明确且具体的改进建议；评估各维度评审反馈是否有援引作文中的具体例子给出具体的评论（而非笼统的意见）；评估各维度评审反馈是否用得体的缓和策略，如使用模糊语，在保持明确性的同时降低批评的威胁性；评估各维度评审反馈是否给出可操作的修改路径。

- 5分 | 卓越：**语气真诚、尊重且鼓励，能先肯定亮点再指出问题；评论紧贴原文，能精确定位到句/词并简要引述或标明位置；缓和自然到位（如“我建议/或许可以/是否考虑...”等），既降低威胁又不模糊结论；给出清晰可执行的修订路径，通常带原因与小示例，必要时还标出优先级（如先改称谓与结尾行动呼吁，再优化句长）。
- 4分 | 较强：**整体积极并兼顾建设性，偶有直白的点评但不刺耳；能较多地对接原文细节并作大致定位，少数地方仍略显概括；缓和策略使用得体但种类偏少；修订建议大都实际可做，但有时缺少示例或理由说明，优先级不够明确。
- 3分 | 合格：**语气中性友好，肯定与指出问题并存但平衡感一般；至少能就一两处给出明确文本定位，其余多为宏观描述；缓和策略偶尔出现但较单一；修订建议给出了基本方向，然而缺少实例化说明或因果解释，学生仍需二次理解才能落实。
- 2分 | 较差：**语气明显消极或带命令感，缺乏对努力点或亮点的认可；几乎不结合文本，难以找到问题所在；没有有效的缓和策略；基本不给修订办法，或给出模糊甚至误导性的指令，让学生无从下手。
- 1分 | 不合格：**语气带攻击或羞辱倾向，易伤害学习动机；完全脱离文本事实或出现明显错误判断；既无缓和也无建设性内容；不给任何可执行建议，整体反馈对学习过程产生负面价值。

任务：

- 1.请你判断以上反馈结果是否合理（合理为 1，不合理为 0）：
- 2.请你判断以上反馈结果的客观证据覆盖度（1-5 分）：
- 3.请你判断以上反馈结果的内部一致性（1-5 分）：
- 4.请你判断以上反馈结果的修改意见质量（1-5 分）：

**English version**

Please now take the role of an evaluator of AWE feedback. Your task is to assess whether a given rater’s score and review feedback are reasonable according to the following criteria. Specifically, you need to make your judgment by considering the essay’s original human-assigned level, its linguistic indicator results, and the objective scoring standards, from the following three perspectives:

**Perspective 1 [Coverage of Objective Evidence]:** Assess whether the review feedback for each dimension makes use of verifiable evidence and whether the statistical statements are accurate.

**5 points | Complete and Verifiable:** The feedback explicitly cites relevant linguistic indicators and provides concrete numerical values. The direction of comparison is accurate, using expressions such as “higher/lower/much higher/much lower,” and it incorporates the scoring standards and provides baselines. When textual evidence is required, specific sentences or passages are identified. The conclusions correspond directly to the evidence, and the analysis can be recalculated and reproduced.

**4 points | Sufficient Evidence, Minor Omissions:** The feedback includes complete indicators and numerical values, and the direction of comparison is correct. However, it either lacks scoring standards/baselines or misses textual references. These omissions do not affect the ability to verify the main conclusions.

**3 points | Partial Evidence:** The feedback mentions indicators but lacks numerical values, or only uses relative terms such as “somewhat higher/somewhat lower.” Alternatively, it may provide numerical values but without baselines while still using comparative wording. The overall direction is generally correct, and no major statistical errors are present.

**2 points | Weak Evidence / Questionable Statistics:** The feedback provides only general judgments or relies on a single indicator to support complex conclusions. It may use exaggerated wording without any thresholds, or it contains at least one major directional error (e.g., reversing higher/lower).

**1 point | Minimal Evidence:** The feedback consists purely of subjective judgment without any supporting basis. The conclusions are clearly inconsistent with the data, or baselines are fabricated. Even when numerical values are provided, their sources are unclear, definitions are ambiguous, or there are multiple statistical inaccuracies, making verification impossible.

**Perspective 2 [Internal Consistency]:** Evaluate the internal logical coherence, consistency in terminology, and alignment between evidence and conclusions within each dimension of the review feedback, as well as across the integrated (overall) review feedback.

**5 points | Fully Consistent:** The arguments support one another without contradictions or misattributions.

**4 points | Largely Consistent:** Occasional minor exaggerations or vague expressions occur, but the core logic remains unaffected; terminology and scope are generally consistent.

**3 points | Partially Inconsistent, Correctable:** A few identifiable inconsistencies are present (e.g., at one point stating “very few conjunctions” and elsewhere “none at all”), but the main conclusion still holds.

**2 points | Multiple Inconsistencies / Leaps in Reasoning:** Several contradictions are evident, or key logical steps are skipped without sufficient supporting evidence.

**1 point | Severely Contradictory:** The logic is irreparable, with conclusions directly negating each other; or conflicting statements within the same passage render the analysis as a whole unintelligible.

**Perspective 3 [Overall Feedback Quality]:** Evaluate whether the review feedback in each dimension is expressed in a tone that is positive, respectful, and highly constructive, combining genuine positive acknowledgment with clear and specific suggestions for improvement. Assess whether the feedback in each dimension draws on concrete examples from the essay to provide specific commentary (rather than vague or generalized remarks). Evaluate whether the feedback employs appropriate mitigation strategies, such as the use of hedging, to reduce the threat to the writer’s face posed by criticism while maintaining clarity. Finally, assess whether the feedback provides actionable pathways for revision.

**5 points | Excellent:** The tone is sincere, respectful, and encouraging, with strengths acknowledged before problems are pointed out. The comments are closely tied to the original text, precisely locating sentences or words with brief quotations or references. Mitigation strategies are naturally and appropriately applied (e.g., “I suggest...,” “perhaps you could...,” “have you considered...”), which reduce the threatening force of criticism without obscuring the conclusions. Clear and actionable revision paths are provided, usually with explanations and small examples, and priorities are indicated when necessary (e.g., first revising forms of address and closing calls to action, then optimizing sentence length).

**4 points | Strong:** The overall tone is positive and constructive, with occasional straightforward comments that are not harsh. The feedback engages with many details from the original text and provides approximate localization, though in some places it remains somewhat general. Mitigation strategies are used appropriately but with limited variety. Most revision suggestions are practical and feasible, but they sometimes lack examples or explanatory reasoning, and the prioritization is not sufficiently clear.

**3 points | Adequate:** The tone is neutral and friendly, balancing praise with critique, though the sense of balance is moderate. At least one or two instances of specific textual localization are provided, while most other comments remain broad or general. Mitigation strategies appear occasionally but are limited in variety. Revision suggestions indicate basic directions but lack concrete examples or causal explanations, requiring students to interpret further before they can implement the feedback.

**2 points | Weak:** The tone is noticeably negative or commanding, with little to no recognition of effort or strengths. The feedback is barely connected to the text, making it difficult to identify specific issues. No effective mitigation strategies are employed. Few or no revision suggestions are given, or the advice provided is vague or even misleading, leaving students without a clear path forward.

**1 point | Inadequate:** The tone is aggressive or humiliating, likely to damage learner motivation. The feedback is completely detached from the textual evidence or contains obvious misjudgments. Neither mitigation nor constructive content is present. No actionable suggestions are offered, and the overall feedback brings negative value to the learning process.

**Task:**

1. Please determine whether the above feedback result is reasonable (1 for reasonable, 0 for not reasonable).
2. Please determine the coverage of objective evidence in the above feedback result (1–5 points).
3. Please determine the internal consistency of the above feedback result (1–5 points).
4. Please determine the quality of the revision suggestions in the above feedback result (1–5 points).

**Appendix I. Evaluation feedback case study**

**Chinese version**

提供的信息:

[作文题目]: 吸烟对个人健康和公众利益的影响

[原始等级]: Intermediate level (B)

[作文内容]: 当今社会，吸烟者有增无减。不少青少年或男女都成吸烟者。吸烟不但个对个人健康破坏也造成公众利益的不好的影响。据科学报告，一根烟里存在二千种以上的化学物品如尼古打等等。这些化学物能导致不少病在人体上。这些病都会让人失去性命以致家破人亡，还要化了不少医药费来医它。除此以为，吸烟者的身体免疫力都会大大减底了。吸烟者都会出现眼红，手颤抖，没精神等状况如没吸烟。这些都会影响个人生活以及工作效率。随手一丢未灭的烟蒂都可能引起火灾。火灾会导致龙大的财物损失。烟不是一个必需品；只是消费品。烟不但有害而且造成了财物损失。如果把钱财花在烟上拿来投资其他方面，显是后者更有用。人们常说，吸烟就如烧了一间屋子。

[作文的语言学指标结果]: 一、词汇维度: NDW: 121、Uber: 78.7645、VV: 0.3511、VV1: 0.7174、SVV1: 23.6739、CVV1: 3.4405、FWR\_conjunction: 0.0533、熵: 6.684、词性熵: 2.9725、MATTR-50: 0.8536、Type ratio of LFP\_Low-level (LFPtr\_L): 0.5702、LFPtr\_M: 0.2231、LFPtr\_H: 0.0744、LD: 0.6036、CCC: 0.0369、MAoAS: 10.2445；二、句法维度: MLS: 18.0667、MLC: 11.2917、NCPS: 1.6、TOTAL\_RTTR: 8.1276、UNIQUE\_RTTR: 2.6056、VO\_RTTR: 3.6056、SP\_RTTR: 4.0、AN\_RTTR: 1.0、AP\_RTTR: 5.0037、CN\_RTTR: 2.0、PP\_RTTR: 0.7071、PV\_RTTR: 1.4142、PC\_RTTR: 2.2361、VO\_RATIO: 0.1857、SP\_RATIO: 0.2286、AN\_RATIO: 0.0143、AP\_RATIO: 0.3857、PP\_RATIO: 0.0286、PV\_RATIO: 0.0286、PC\_RATIO: 0.0714、MDD: 2.1006、NTCU\_RATIO: 0.4、TCMDD: 1.8984、MLTTCU: 12.6667；三、语篇衔接维度: pron\_density: 0.0355、pron\_num\_per\_sent: 0.4、per-pron\_diversity: 0.0476、subj\_density: 0.1065、subj\_pron\_diversity: 1.0、conj\_density: 0.0533、conj\_num\_per\_sent: 0.6；四、语言正确度维度: norm\_total\_errors: 12.426、norm\_Chi\_character\_errors: 4.7337、norm\_Chi\_word\_errors: 5.3254、norm\_sentence\_errors: 2.3669、norm\_discourse\_errors: 0.0

模型的反馈:

1. 评分: Advanced level (A)
2. 评审意见-内容回应与交际功能:

该作文紧扣题目“吸烟对个人健康和公众利益的影响”，全面覆盖了个人健康损害（如疾病、免疫力下降、生理症状）、公共安全风险（火灾、财物损失）以及经济浪费（投资对比）三个核心维度，内容充实、逻辑清晰，具备较强的交际意图。作者通过“科学报告”“医药费”“投资其他方面”等具体例证增强说服力，结尾用“吸烟就如烧了一间屋子”这一比喻性总结，有效提升了情感感染力与思想深度。从交际功能看，全文实现了从现象描述到价值判断的递进，符合高级作文“内容充实、表达准确恰当”的要求。然而，部分表述略显口语化（如“没精神等状况如没吸烟”），削弱了正式性；且“家破人亡”“龙大的财物损失”等表达虽具冲击力，但略显夸张，若能更精准使用“家庭破裂”“重大财产损失”等规范表达，将更符合学术写作语境。总体而言，内容回应完整、意图明确，符合 A 档“内容充实、表达准确恰当”的标准，但语言表达的精确性仍有提升空间。

3. 评审意见-词汇:

该作文在词汇维度表现优异，多项指标均显著优于 A 档平均水平。NDW（121）虽低于 A 档均值（149.041），但 Uber 指数高达 78.7645，远超 A 档均值（59.639），表明词汇重复率极低，词汇丰富；LFPtr\_L（0.5702）显著低于 A 档均值（0.611），说明作者有效减少了基础词汇依赖；LFPtr\_M（0.2231）和 LFPtr\_H（0.0744）均高于 A 档均值（0.156 和 0.059），显示中高级词汇使用比例显著提升，如‘化学物品’‘免疫力’‘投资’‘消费’‘火灾’等词使用精准；MAoAS（10.2445）远超 A 档均值（9.644），表明作者大量使用了习得年龄较晚、抽象度高的词汇（如“破坏”“导致”“造成”“损失”“浪费”），词汇成熟度极高；LD（0.6036）显著高于 A 档均值（0.466），说明信息密度高，内容承载丰富；CCC（0.0369）高于 A 档均值（0.032），显示使用了较多复杂汉字。然而，词汇使用存在误用问题，如“尼古打”应为“尼古丁”，“家破人亡”语义过重，“龙大的财物损失”搭配不当（应为“重大”），说明词汇掌握虽广，但语义精确性不足。建议：在保持词汇丰富性的基础上，加强语义搭配训练，尤其注意量词、专业术语的规范使用。

4. 评审意见-句法:

该作文句法表现呈现高复杂度但低多样性的矛盾特征。MLS（18.0667）显著低于 A 档均值（32.991），表明句子整体偏短，缺乏长句的复杂结构；NCPS（1.6）远低于 A 档均值（3.051），说明作者极少使用流水句或多分句结构，句式结构单一，未能充分体现汉语意合特征；MLC（11.2917）接近 A 档均值（11.116），但因 NCPS 低，句子虽短却缺乏层次，语义承载集中，显得生硬。句法多样性方面，TOTAL\_RTTR（8.1276）和 UNIQUE\_RTTR（2.6056）

均高于 A 档均值（8.513 和 2.542），说明搭配多样性良好，尤其在汉语特有搭配（CN\_RTTR=2.0, PC\_RTTR=2.2361）上表现突出，如“一根烟”等量名使用准确。但 VO\_RATIO（0.1857）和 SP\_RATIO（0.2286）低于 A 档均值（0.209 和 0.183），说明主谓宾结构使用比例偏低，符合一定的汉语话题突出特征，但因 NCPS 低，未能形成有效话题链。MDD（2.1006）低于 A 档均值（2.969），表明句法结构依赖简单线性排列，缺乏深层依存关系。话题链句数比例（0.4）高于 A 档均值（0.241），TCMDD（1.8984）和 MLTTCU（12.6667）也高于 A 档均值，说明作者在局部段落中已能构建较完整的话题链，如“吸烟者→身体免疫力减底→眼红、手颤抖→没精神→影响生活→工作效率”，这是本作文最突出的句法优势。建议：在保持话题链优势的基础上，增加 NCPS，尝试将多个短句通过“也”“还”“而且”等连接词串联为流水句，如“吸烟者不仅身体免疫力下降，还常出现眼红、手颤抖、没精神等状况，这些都会影响他们的生活质量和工作效率”，以提升句法复杂度与自然度。

5. 评审意见-语篇衔接:

该作文在语篇衔接上表现出意合特征初步显现但未完全成熟的特征。连词密度（0.0533）与 A 档均值（0.042）持平，句均连词数（0.6）略高于 A 档均值（0.841），说明作者虽未过度依赖显性连接词，但连词使用频率仍偏低，未能充分支撑复杂逻辑关系（如因果、递进、转折）。主语密度（0.1065）略高于 A 档均值（0.099），说明仍存在少量显性主语重复，如“吸烟者”多次出现，若能通过话题链自然承接，可进一步提升衔接效率。代词主语多样性（1.0）异常偏高，说明作者在极少数句子中使用了“他”“他们”等代词，但使用频率极低，缺乏变化，未能体现多样性。人称代词多样性（0.0476）远低于 A 档均值（0.177），表明代词使用单一，仅限于“吸烟者”“人们”。整体而言，语篇衔接以零照应为主，符合汉语意合特征，但缺乏代词与连接词的灵活调度。建议：在保持零照应优势的基础上，适度增加“因此”“此外”等连词，或在话题链中使用“这”“此”“如此”等指示代词，以增强语篇连贯性，使“意合”更自然、更丰富。

6. 评审意见-语言正确度:

该作文语言正确度整体处于 A 档边缘，标准化偏误总数（12.426）略高于 A 档均值（10.920），但远低于 B 档（10.979）和 C 档（14.175），说明语言错误控制良好。字偏误（4.7337）和词偏误（5.3254）是主要问题，集中于词汇误用：如“尼古打”应为“尼古丁”，“龙大的财物损失”应为“重大”，“家破人亡”语义过重，“显是后者更有用”应为“显然后者更有用”。句偏误（2.3669）主要表现为结构残缺与搭配错误，如“这些病都会让人失去性命以致家破人亡”中“以致”使用不当（应为“进而”或“导致”），“随手一丢未灭的烟蒂都可能引起火灾”缺少主语（“吸烟者”）。篇章偏误为 0，说明段落结构完整。总体而言，语言错误数量控制在高级水平，但高阶词误用与搭配不当是主要瓶颈，影响表达的准确性与专业性。建议：重点修改高频误用词，建立“搭配库：（如“导致……”“造成……”“引发……”），并加强语义一致性训练，使语言在丰富性与准确性之间实现平衡。

English version

[Essay Title]: The Impact of Smoking on Personal Health and Public Interests

[Original Level]: Intermediate level (B)

[Essay Content]: 当今社会，吸烟者有增无减。不少青少年或男女都成吸烟者。吸烟不但个对个人健康破坏也造成公众利益的不好的影响。据科学报告，一根烟里存在二千种以上的化学物品如尼古打等等。这些化学物能导致不少病在人体上。这些病都会让人失去性命以致家破人亡，还要化了不少医药费来医它。除此以为，吸烟者的身体免疫力都会大大减底了。吸烟者都会出现眼红，手颤抖，没精神等状况如没吸烟。这些都会影响个人生活以及工作效率。随手一丢未灭的烟蒂都可能引起火灾。火灾会导致龙大的财物损失。烟不是一个必需品；只是消费品。烟不但有害而且造成了财物损失。如果把钱财花在烟上拿来投资其他方面，显是后者更有用。人们常说，吸烟就如烧了一间屋子。

[Linguistic Index Results of the Essay]

1. Lexical Dimension: NDW: 121; Uber Index: 78.7645; VV: 0.3511; VV1: 0.7174; SVV1: 23.6739; CVV1: 3.4405; FWR\_conjunction: 0.0533; Entropy: 6.684; POS Entropy: 2.9725; MATTR-50: 0.8536; Type ratio of LFP\_Low-level (LFPtr\_L): 0.5702; LFPtr\_M: 0.2231; LFPtr\_H: 0.0744; LD: 0.6036; CCC: 0.0369; MAoAS: 10.2445

2. Syntactic Dimension:

MLS: 18.0667; MLC: 11.2917; NCPS: 1.6; TOTAL\_RTTR: 8.1276; UNIQUE\_RTTR: 2.6056; VO\_RTTR: 3.6056; SP\_RTTR: 4.0; AN\_RTTR: 1.0; AP\_RTTR: 5.0037; CN\_RTTR: 2.0; PP\_RTTR: 0.7071; PV\_RTTR: 1.4142; PC\_RTTR: 2.2361; VO\_RATIO: 0.1857; SP\_RATIO: 0.2286; AN\_RATIO: 0.0143; AP\_RATIO: 0.3857; PP\_RATIO: 0.0286; PV\_RATIO: 0.0286; PC\_RATIO: 0.0714; MDD: 2.1006; NTCU\_RATIO: 0.4; TCMDD: 1.8984; MLTTCU: 12.6667.

3. Discourse Cohesion Dimension:

Pronoun density: 0.0355; Pronouns per sentence: 0.4; Personal pronoun diversity: 0.0476; Subject density: 0.1065; Subject pronoun diversity: 1.0; Conjunction density: 0.0533; Conjunctions per sentence: 0.6.

4. Language Accuracy Dimension:

Standardized total errors: 12.426; Standardized Chinese character errors: 4.7337; Standardized Chinese word errors: 5.3254; Standardized sentence errors: 2.3669; Standardized discourse errors: 0.0.

Model Feedback

1. Score: Advanced level (A)

2. Comments – Content Response and Communicative Function:

The essay closely follows the topic “吸烟对个人健康和公众利益的影响” (“The impact of smoking on personal health and public welfare”), and it thoroughly addresses three core dimensions: personal health damage (such as diseases, weakened immunity, and physiological symptoms), public safety risks (such as 火灾 “fire” and 财物损失 “property loss”), and economic waste (such as 投资对比 “investment comparison”). The content is substantial, logically coherent, and demonstrates

a clear communicative purpose. Persuasiveness is strengthened through concrete examples like “科学报告” (“scientific reports”), “医药费” (“medical expenses”), and “投资其他方面” (“investment in other areas”). The conclusion uses the metaphor “吸烟就如烧了一间屋子” (“smoking is like burning down a house”), which enhances both emotional appeal and intellectual depth. From a communicative perspective, the essay progresses from describing phenomena to making value judgments, aligning with the advanced-level requirement of “substantial content and accurate, appropriate expression.” Some expressions, however, are somewhat colloquial. For instance, “没精神等状况如没吸烟” (“conditions like lack of energy when not smoking”) reduces the level of formality. In addition, phrases such as “家破人亡” (“family ruined and people dead”) and “龙大的财物损失” (“enormous property loss”) are rhetorically strong but slightly exaggerated. More precise alternatives like “家庭破裂” (“family breakdown”) and “重大财产损失” (“significant property loss”) would be more appropriate in the context of academic writing. Overall, the essay provides a complete and well-focused response with a clear communicative intent, meeting the A-level standard of “substantial content and accurate, appropriate expression.” Nevertheless, there remains room for improvement in linguistic precision.

### 3. Comments – Lexical Dimension

The essay performs strongly in the lexical dimension, with multiple indices significantly exceeding the advanced-level averages. Although NDW (121) is lower than the A-level mean (149.041), the Uber index reaches 78.7645, well above the mean (59.639), indicating a very low word repetition rate and mature lexical deployment. LFPtr\_L (0.5702) is notably lower than the A-level mean (0.611), suggesting reduced reliance on basic vocabulary. Both LFPtr\_M (0.2231) and LFPtr\_H (0.0744) are higher than the A-level means (0.156 and 0.059), reflecting a marked increase in the use of mid- and high-level words, such as 化学物品 (“chemical substances”), 免疫力 (“immunity”), 投资 (“investment”), 消费 (“consumption”), and 火灾 (“fire”). MAoAS (10.2445) is also well above the A-level mean (9.644), indicating frequent use of words acquired later and with higher levels of abstraction, such as 破坏 (“to destroy”), 导致 (“to lead to”), 造成 (“to cause”), 损失 (“loss”), and 浪费 (“waste”), suggesting very high lexical maturity. LD (0.6036) is significantly higher than the A-level mean (0.466), pointing to high information density and rich content load. CCC (0.0369) exceeds the A-level mean (0.032), showing increased use of complex characters. However, there are instances of lexical misuse. For example, 尼古打 was incorrectly used instead of 尼古丁 (“nicotine”); the phrase 家破人亡 (“family ruined and people dead”) is overly heavy in meaning; and 龙大的财物损失 is semantically inappropriate, where 重大 (“significant/major”) should have been used instead. These issues suggest that although lexical breadth is strong, semantic precision is insufficient. Recommendation: While maintaining lexical richness, greater attention should be paid to collocational accuracy, especially in the use of classifiers and technical terms.

### 4. Comments – Syntactic Dimension

The essay demonstrates a paradoxical syntactic profile characterized by high complexity but low diversity. MLS (18.0667) is far below the A-level mean (32.991), indicating that sentences are generally short and lack the complex structures typical of longer sentences. NCPS (1.6) is also much lower than the A-level mean (3.051), suggesting that the author rarely employs run-on or multi-clause sentences, resulting in a monotonous sentence structure that fails to fully reflect the paratactic feature of Chinese. MLC (11.2917) is close to the A-level mean (11.116), but given the low NCPS, the sentences—though short—lack internal layering, leading to dense semantic loading and a somewhat stiff expression. In terms of syntactic diversity, both TOTAL\_RTTR (8.1276) and UNIQUE\_RTTR (2.6056) exceed the A-level means (8.513 and 2.542), reflecting good collocational variety, particularly in Chinese-specific combinations (CN\_RTTR = 2.0, PC\_RTTR = 2.2361). For instance, the accurate use of “一根烟” (“a cigarette,” with correct classifier–noun pairing) exemplifies appropriate collocational competence. However, VO\_RATIO (0.1857) and SP\_RATIO (0.2286) are lower than the A-level means (0.209 and 0.183), suggesting underuse of subject–predicate–object structures. While this partly aligns with Chinese topic-prominent tendencies, the low NCPS prevents the effective construction of topic chains. MDD (2.1006) falls below the A-level mean (2.969), indicating reliance on simple linear arrangements with insufficient deep dependency relations. In contrast, the proportion of topic-chain sentences (0.4) is higher than the A-level mean (0.241), and both TCMDD (1.8984) and MLTTCU (12.6667) also exceed A-level means, showing that the author is capable of constructing relatively complete topic chains at the paragraph level. For example: “吸烟者 → 身体免疫力降低 → 眼红、手颤抖 → 没精神 → 影响生活 → 工作效率” (“smoker → weakened immunity → red eyes, hand tremors → lack of energy → affects daily life → reduces work efficiency”). This constitutes the essay’s most notable syntactic strength. Recommendation: While maintaining the strength in topic-chain construction, the author should increase NCPS by linking short sentences into run-on structures using conjunctions such as “也” (“also”), “还” (“still”), and “而且” (“moreover”). For example: “吸烟者不仅身体免疫力下降，还常出现眼红、手颤抖、没精神等状况，这些都会影响他们的生活质量和工作效率” (“Smokers not only experience decreased immunity but also often suffer from red eyes, hand tremors, and lack of energy, all of which affect their quality of life and work efficiency”). This would enhance both syntactic complexity and naturalness.

### 5. Comments – Discourse Cohesion

The essay demonstrates emerging but not yet fully mature features of paratactic cohesion. Conjunction density (0.0533) is close to the A-level mean (0.042), while the average number of conjunctions per sentence (0.6) is slightly higher than the A-level mean (0.841). This indicates that although the author does not overly rely on explicit connectives, the frequency of conjunction use remains relatively low and fails to adequately support more complex logical relations (e.g., causality, progression, contrast). Subject density (0.1065) is marginally above the A-level mean (0.099), suggesting some repetition of explicit subjects, such as the repeated use of “吸烟者” (“smokers”). Greater reliance on topic chains could enhance cohesion efficiency. The diversity of pronominal subjects (1.0) is unusually high, showing that the author occasionally uses pronouns such as “他” (“he”) and “他们” (“they”), but these appear too infrequently and with little variation to establish



genuine diversity. The diversity of personal pronouns (0.0476) is far below the A-level mean (0.177), indicating that pronominal use is highly limited, largely restricted to “吸烟者” (“smokers”) and “人们” (“people”). Overall, cohesion is primarily achieved through zero anaphora, which is consistent with the paratactic nature of Chinese. However, the essay lacks flexibility in deploying pronouns and conjunctions. Recommendation: While retaining the advantage of zero anaphora, the author should moderately increase the use of conjunctions such as “因此” (“therefore”) and “此外” (“in addition”), or employ demonstrative pronouns such as “这” (“this”), “此” (“this”), and “如此” (“thus”) within topic chains to enhance discourse coherence. This would render the parataxis more natural and more nuanced.

#### 6. Comments – Linguistic Accuracy

The overall linguistic accuracy of this essay is at the borderline of the A level. The standardized total number of errors (12.426) is slightly higher than the A-level mean (10.920), but much lower than the B-level (10.979) and C-level (14.175), indicating good control of language errors. Character errors (4.7337) and word errors (5.3254) are the main problems, concentrated in lexical misuse. For example, “尼古打” should be “尼古丁” (nicotine); “龙大的财物损失” should be “重大” (major); “家破人亡” is semantically too strong; and “显是后者更有用” should be “显然后者更有用” (obviously the latter is more useful). Sentence errors (2.3669) are mainly reflected in incomplete structures and collocational mistakes. For instance, in “这些病都会让人失去性命以致家破人亡” the use of “以致” is inappropriate (it should be “进而” or “导致”), and “随手一丢未灭的烟蒂都可能引起火灾” lacks a subject (which should be “吸烟者” “smoker”). At the discourse level, the error count is 0, indicating that the paragraph structure is complete. Overall, the number of language errors is controlled at an advanced level. However, the misuse of advanced words and improper collocations are the main bottlenecks, affecting the accuracy and professionalism of expression. Suggestion: Focus on correcting frequently misused words, establish a “collocation bank” (e.g., “导致……” lead to..., “造成……” cause..., “引发……” trigger...), and strengthen training on semantic consistency so that a balance between richness and accuracy can be achieved.