Beyond a Writer Label: Rich Context Shapes AI-Text Detector Scores

Anonymous authors
Paper under double-blind review

Abstract

AI-text detector evaluations depend on how test passages are prompted. We ask whether a brief writer identity captures the response to a richly developed background. Across 150 Chinese literary-writing topics and five generators (4,500 texts), a one-sentence writer label changes detector scores little, while a detailed fictional background produces much larger shifts. The background also lowers scores more than the tested anti-AI-style directive: by 0.062 on AIGC-ZHv2's classifier scale and 0.098 corpus-rank units for Fast-DetectGPT. Two further biographies preserve the strategy-level separation. In a separate three-generator cohort, an independent second draw preserves the aggregate score-shift directions, and life-history facts carry the larger component response within the complete persona. A detailed description of a fictional town, with no assigned writer identity, then produces a reduction similar to that life-fact bundle across the 12 statistical outputs (median 0.264 versus 0.257 within the same experiment). Rich contextual information can therefore elicit large detector-score shifts without an explicit writer identity. Evaluations should distinguish brief labels from developed contexts and report their generator-specific responses.

1 Introduction

An AI-text detector is evaluated on passages produced under particular generation prompts. Those prompts can offer very different amounts of material for writing: one may say only "You are an experienced writer," while another supplies a writer's family, jobs, relationships, and places. Treating both as persona prompting hides a consequential choice in the evaluation design. The question is whether naming a writer elicits the same detector response as giving the generator a developed context.

Prior work shows that task constraints affect detection (Koike et al., 2024), author attributes influence detector behavior (Li & Wan, 2025), and personalized imitation can alter detection features (Gao et al., 2026). We compare task-only generation (N), a one-sentence writer label (S), a detailed fictional writer background (P1), and a direct anti-AI-style directive (A) within the same topic–generator cells. We then test two further biographies, separate P1's life facts from its narrative and temperament cues, and independently sample the central contrasts. A final control replaces the life facts with a developed description of a fictional town (C), without assigning a writer identity.

The label produces small score changes; the developed writer background produces much larger ones and also lowers scores more than the tested directive. Within the complete persona, life-history facts carry the larger component response. The town context reproduces that bundle's score reduction in its own experiment. Together, these findings show that *rich contextual information can drive detector-score shifts without an explicit writer identity*. Detector evaluations should record and vary the content supplied to generators, rather than treating a role label as representative of developed contexts.

2 Comparing prompt strategies

For a topic t, generator g, prompt condition c, and fixed recorded decoding settings, let $x_{t,g,c}$ be the generated passage. We compare a fixed scoring map a_d across conditions. Its sign follows the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

implementation's designated AI-oriented direction. Our main estimand is the within-topic difference in detector scores between prompting strategies.

For each output, our observed paired contrast is

$$\widehat{\Delta}_{d,t}(c,c') = \frac{1}{G}\sum_{g=1}^{G}\left[a_d(x_{t,g,c}) - a_d(x_{t,g,c'})\right], \qquad \widehat{\Delta}_d(c,c') = \frac{1}{T}\sum_{t=1}^{T}\widehat{\Delta}_{d,t}(c,c'). \tag{1}$$

The topic is the inference unit. We average generators within each topic and also report generator-specific contrasts. Experiments I–III and V contain one passage per topic–generator–condition cell; Experiment IV-B contains two draws and averages them within that cell before topic-level inference. Main-text baseline contrasts set $(c, c') = (b, P_j)$ in Equation 1, so a positive difference indicates a lower score under the persona; the archived contrast files retain their original signs.

For summary analyses of the three fixed personas, we average the P1, P2, and P3 scores within each topic–generator cell; individual contrasts retain their differences.

For the construct-validity control, $R_d = \widehat{\Delta}_d(N, C)/\widehat{\Delta}_d(N, B)$ expresses the non-biographical context's reduction as a fraction of the life-fact bundle's within the same scoring corpus. We bootstrap both legs on shared topic indices.

Score interpretation. For the 12 MGT-Eval statistics, we use the saved empirical mid-ranks with the implementation-defined signs,

$$a_d(x_i) = \frac{\text{rank}_{\text{avg}}(s_d(x_i); \{s_d(x_j)\}_{j=1}^{n_d})}{n_d}, \tag{2}$$

where s_d incorporates the implementation's sign convention. These scores measure relative position in a common scoring corpus for all conditions within an experiment. Its reference comprises 4,500 machine passages and 150 human excerpts in Experiment I, and machine passages alone in Experiments II–V (270, 480, 360, 540, and 450 respectively). We hold each reference fixed for its experiment and compare rank-scale magnitudes only within that experiment. Classifiers and the auxiliary style score retain their own scales. Our outcome is the paired detector-score response to prompting, rather than a fixed-threshold detection decision.

3 Study design

Table 1: Complementary designs and their inferential roles. Experiments II and III reuse 90 and 360 main-study texts; IV-B reuses 180 IV-A texts. V independently regenerates its conditions on the same 30 topics and three generators as IV-A. The study contains 5,970 distinct generation records.

Design	Topics	Models	Texts	Question addressed
I: Strategies	150	5	4,500	Background, style instruction, role label
II: Components	45	2	270	Role of personality fragments
III: Personas	30	2	480	Variation across biographies
IV-A: Decomposition	30	3	360	Life facts versus narrative/style cues
IV-B: Resampling	30	3	540	Independent-draw robustness
V: Town context	30	3	450	Writer life facts versus detailed setting

3.1 Generation and prompt interventions

Experiment I: complete strategies. We use 150 literary-writing topics in five categories, with 30 topics per category: memory and identity, relationships, urban everyday life, nature and seasons, and uncertainty and transition. The operative list was frozen for this study; the supplementary file `data/topics/paper_topics.csv` releases every category label and Chinese topic string. The generators are DeepSeek-V3, GLM-4.5-Air, Kimi-K2-Instruct, MiniMax-M2.1, and Qwen3-32B. All calls use DashScope, temperature 0.9, top-p 0.9, a 900-token output cap, and an empty system prompt. Every condition requests approximately 500 Chinese characters of plain literary text, without a title or process explanation. Appendix A records the provider identifiers, timestamps, and corpus provenance.

The six conditions establish interpretable reference points. N supplies the writing task alone. S adds only "You are an experienced Chinese-language writer." P1 develops a fictional writer through family, education, work, relationships, temperament, and observational habits. A discourages generic and mechanical expression and adds factuality and rhetorical instructions; P1+A combines the two. R uses public biographical information about a real writer. P1 and A differ by 16 prompt characters for the same topic, so their comparison tests two similarly long, recorded strategies rather than isolated prompt attributes. The exact Chinese prompts appear in Appendix E.

Experiment II: a component intervention. On 45 topics and two generators, DeepSeek and Kimi, PN replaces three explicit personality and observational fragments in P1 with character-matched daily-life facts; PR removes those fragments. Both variants retain the surrounding biography. This local component check reuses 90 P1 passages from Experiment I. Appendix B.4 gives the edits, token counts, and full results.

Experiment III: distinct persona instances. On 30 topics balanced across the five categories and the same two generators, we compare N, S, A, P1, P1+A, R, P2, and P3. P2 describes a coastal freelance writer and community-radio editor, emphasizing colloquial speech, action, and dialogue. P3 describes an archivist in northwestern China, emphasizing objects, local records, and restrained observation. The three personas are fixed AI-generated descriptions. P2 is 10 characters longer than P1; P3 is 69 characters shorter. The corpus reuses 360 passages for the original six conditions and adds 120 P2/P3 passages. The follow-ups were collected after the main batch with the same recorded provider identifiers and decoding settings.

Experiment IV: life facts, narrative cues, and new draws. On the 30 Experiment III topics, a separate three-generator cohort—DeepSeek-V3.1, Kimi-K2.5, and GLM-4.5—uses the same decoding settings in non-reasoning mode. IV-A compares N, P1, and two deterministic subsets of P1. B retains life-history facts and deletes temperament, observational preferences, thematic orientations, and narrative-aesthetic cues; T retains those cues without life events. Both are assembled by literal deletion and minimal grammatical repair, not by model rewriting. Thus P1 is the recorded B+T bundle. We estimate

$$\Delta_B = \tfrac{1}{2}[(N - B) + (T - P1)], \quad \Delta_T = \tfrac{1}{2}[(N - T) + (B - P1)], \quad I_{BT} = B + T - N - P1,$$

where positive Δ_B or Δ_T denotes a score reduction and positive I_{BT} denotes an additional reduction beyond additivity.

IV-B evaluates sampling robustness for N, A, and P1. Draw 1 reuses IV-A's N and P1 and adds A; draw 2 independently regenerates all three conditions. The two draws are nested within each topic–generator–condition cell and are not treated as additional topics. The provider does not expose a fixed sampling seed, so this is independent resampling under matched decoding settings rather than seeded reproduction.

Experiment V: a detailed setting without a writer. Experiment V tests whether a developed context can produce B's score response without specifying a writer. On the same 30 topics and three-generator cohort as IV-A, it independently generates N, B, A, a non-biographical context C, and a directive variant A′. The B template is identical to IV-A's, but all V passages are newly generated. C describes a fictional Yangtze-river town through its population, trades, landmarks, produce, and weather; it assigns no writer identity or personal history. Designed as another detailed, concrete context, C is 73 template characters shorter than B; their generated passages have nearly identical mean lengths (655.5 versus 653.8 characters). Complete prompts and the corpus audit appear in Appendices E and B.7.

3.2 Detection outputs

The scoring panel contains 24 outputs: five long-text aggregations for each of two Chinese classifiers (AIGC-ZHv2 and RoBERTa-ZH), 12 MGT-Eval statistical outputs, a handcrafted style proxy, and an ensemble. Main results retain one aggregation per classifier, giving 16 default outputs. Follow-ups analyze the 12 statistics, style proxy, and ensemble; Experiment V has 13 signals because its batch has no ensemble. Classifier scores saturate near their lower bound in the follow-ups, so their full results and floor diagnostics appear in the appendix.

The statistics cover likelihood, token rank, entropy, GLTR, log-rank ratio, LastDE/LastDE++, conditional curvature, mutation-based scoring, a Binoculars configuration, and TOCSIN (Gehrmann et al., 2019; Su et al., 2023; Bao et al., 2024; Xu et al., 2025; Ma & Wang, 2024; Zhu et al., 2025). The reference language model is `uer/gpt2-chinese-cluecorpussmall`. The implemented Binoculars variant uses identical observer and performer identifiers and is labeled *same-LM* (Hans et al., 2024). The main figures show the other 11 statistical outputs. Several statistics share models or score constructions, so output counts describe responses across the panel rather than independent replications.

The auxiliary ensemble uses supervised logistic stacking with generator-family holdout in Experiment I and a mean of training-family empirical ranks in the machine-only follow-ups. Appendix A details both constructions, normalization, score directions, and aggregation.

3.3 Inference and evidence accounting

We compute 95% percentile intervals from 5,000 topic bootstrap resamples and two-sided Wilcoxon signed-rank tests. Holm correction covers the 24 outputs separately for each Experiment I contrast and the 14 analysis signals—13 for Experiment V—for each follow-up estimand and generator stratum (Holm, 1979). Standardized paired effects are $d_z = \overline{\Delta}/\,\mathrm{sd}(\Delta_t)$. Topics are the inference units; intervals condition on the recorded generations. Experiments IV and V report their analysis families in the main text and retain the ten classifier/window outputs as a supplementary family. Table 1 and Appendix A specify corpus reuse.

4 Results

4.1 A writer label does not reproduce the developed-background response

Table 2: Experiment I: score reductions from task-only prompting (N), averaged over five generators and 150 topics. Positive values indicate lower scores than N. Classifiers retain native scales; Fast-DetectGPT uses corpus ranks. Direct paired S–P1 inference appears in Appendix B.2.

Output	Writer label N–S	Style directive N–A	Background N–P1
AIGC-ZHv2	0.0130	0.1744	0.2365
RoBERTa-ZH	0.0001	-0.0003	0.0003
Fast-DetectGPT	0.0174	0.0466	0.1446

The one-sentence identity S yields small score changes relative to task-only prompting N. AIGC-ZHv2 shifts by 0.013 under S versus 0.237 under P1; Fast-DetectGPT shifts by 0.017 versus 0.145 (Table 2). Across the 16 default outputs, the paired N–S standardized effects range from 0.016 to 0.231.

The direct S–P1 comparison confirms the separation: 0.224 on AIGC-ZHv2 ($d_z = 1.792$) and 0.127 rank units on Fast-DetectGPT ($d_z = 1.417$), both with lower P1 scores. The brief label does not reproduce the magnitude of the developed background's response; the complete paired results appear in Appendix B.2.

4.2 A developed background produces larger score reductions than the tested directive

The complete background P1 also receives lower scores than the recorded anti-AI-style directive A in the topic-paired, five-generator average. The A–P1 difference is 0.062 on AIGC-ZHv2 ($d_z = 0.664$), 0.098 rank units on Fast-DetectGPT ($d_z = 1.051$), and 0.150 on Rank ($d_z = 1.787$). The RoBERTa-ZH difference is much smaller, at 0.0006 on its native scale. Table 3 and Appendix B give the remaining outputs and baseline contrasts.

Table 3: Experiment I: A minus P1, averaging five generators within each of 150 topics. Positive values mean lower P1 scores. All 16 default outputs are shown once; eight alternative classifier aggregations appear in the appendix. Intervals are pointwise 95% topic-bootstrap intervals. All comparisons have Holm $p < 0.001$ under the original 24-output family. Classifier and corpus-rank scales are distinct.

Output	A–P1	95% interval	d_z
Trained classifiers			
AIGC-ZHv2	0.0621	[0.0473, 0.0773]	0.664
RoBERTa-ZH	0.0006	[0.0002, 0.0012]	0.178
Statistical outputs			
Binoculars (same LM)	0.0963	[0.0812, 0.1117]	1.003
DNA-DetectLLM	0.0732	[0.0573, 0.0885]	0.742
Entropy	0.0751	[0.0603, 0.0896]	0.819
Fast-DetectGPT	0.0980	[0.0834, 0.1125]	1.051
GLTR	0.1054	[0.0923, 0.1183]	1.312
LastDE	0.1805	[0.1612, 0.2005]	1.447
LastDE++	0.1029	[0.0878, 0.1183]	1.069
Likelihood	0.1003	[0.0883, 0.1129]	1.284
LogRank	0.0958	[0.0829, 0.1081]	1.223
LRR	0.0740	[0.0591, 0.0897]	0.768
Rank	0.1503	[0.1368, 0.1637]	1.787
TOCSIN	0.0925	[0.0771, 0.1076]	0.961
Auxiliary constructions			
Style proxy	0.0373	[0.0319, 0.0429]	1.056
Ensemble	0.0287	[0.0118, 0.0446]	0.284

4.3 Additional character backgrounds preserve the strategy-level separation

Two further fixed biographies, P2 and P3, preserve the separation from the directive A across the follow-up panel. For Fast-DetectGPT, N–P1, N–P2, and N–P3 are 0.180, 0.355, and 0.281 corpus-rank units, with intervals $[0.099, 0.262]$, $[0.275, 0.436]$, and $[0.217, 0.342]$. P2 produces the largest observed shift in this fixed set (Appendix Figure 2). The brief writer label also scores above the three-background average $\overline{P}$ across the follow-up panel (Appendix Table 10).

4.4 Concrete life-history facts carry the larger component response

Table 4: Experiment IV-A decomposition, averaging three generators within each of 30 topics. Positive values indicate score reduction; Holm significance is computed within the 14-output family for each row. Main effects average the two corresponding simple effects. B–P1 measures T's incremental score response when B is present.

Estimand	Positive	Holm sig.	Median reduction	Median d_z
Biography main effect	14/14	14/14	0.215	2.558
Style main effect	12/14	2/14	0.043	0.351
B–P1	4/14	0/14	-0.018	-0.126
B×T interaction	0/14	3/14	-0.118	-0.482

The life-fact bundle B carries the larger response within the P1 partition: its median main-effect reduction is 0.215 ($d_z = 2.558$), compared with 0.043 ($d_z = 0.351$) for the temperament and narrative-cue bundle T (Table 4). B reduces scores both when added to N and when added to T. Adding T to B, by contrast, produces no consistent further reduction (median B–P1 $= -0.018$). The negative median interaction (-0.118) indicates overlap between the two component responses on this score scale. B is a bundle of life facts and lived settings; Experiment V tests whether the larger response requires those details to describe a writer's life.

Independent draws recover the aggregate directions. Both independent IV-B draws recover positive average N–P1 and A–P1 contrasts across the follow-up panel. Their median N–P1 reductions

are 0.277 and 0.201; their median A–P1 reductions are 0.291 and 0.223 (Appendix B.6). The aggregate response persists across stochastic generations.

4.5 A detailed setting elicits a similar score response

Table 5: Experiment V: score reductions from task-only prompting (N), averaging three generators within each of 30 topics. Positive values indicate lower scores than N. B supplies writer life facts; C supplies a detailed fictional setting without an assigned writer; A is the recorded directive and A′ its recorded variant. $R_d = \widehat{\Delta}_d(N, C)/\widehat{\Delta}_d(N, B)$ with a 95% topic-bootstrap interval. Asterisks mark Holm-adjusted $p < 0.05$ within the 13-signal family for each contrast. Statistical values use Experiment V's own 450-text rank reference.

Output	N–B	N–C	N–A	N–A′	R_d	R_d interval
Likelihood	0.262*	0.265*	0.002	-0.076*	1.01	[0.79, 1.26]
Rank	0.228*	0.253*	-0.084	-0.157*	1.11	[0.88, 1.41]
LogRank	0.254*	0.263*	-0.005	-0.090*	1.04	[0.81, 1.28]
Entropy	0.186*	0.204*	-0.039	-0.121*	1.10	[0.80, 1.44]
GLTR	0.234*	0.254*	-0.028	-0.113*	1.08	[0.86, 1.38]
LRR	0.199*	0.241*	-0.015	-0.117*	1.21	[0.92, 1.61]
LastDE	0.161*	0.222*	-0.093	-0.090	1.38	[1.12, 1.78]
LastDE++	0.280*	0.280*	0.055	0.017	1.00	[0.82, 1.22]
Fast-DetectGPT	0.264*	0.274*	0.065	0.005	1.04	[0.83, 1.30]
DNA-DetectLLM	0.260*	0.273*	0.091	0.028	1.05	[0.83, 1.33]
Binoculars (same-LM)	0.272*	0.276*	0.068	0.008	1.02	[0.82, 1.27]
TOCSIN	0.263*	0.273*	0.059	-0.008	1.04	[0.82, 1.29]

With topics, generators, decoding settings, and the task sentence fixed, C produces a strong score reduction without assigning a writer identity. Across the 12 statistical outputs, its median N–C reduction is 0.264, close to B's 0.257 within the same scoring corpus (Table 5). The median per-output reproduction ratio is 1.05. Both contexts produce large responses for DeepSeek and GLM and small responses for Kimi-K2.5 (Appendix B.7).

This result changes the interpretation of the background effect. The life-fact bundle B is powerful, but a detailed setting without a writer yields a similar detector-score response. The longest prompt, directive A, does not produce the same reduction in this batch. Richly developed context, rather than an explicit writer identity or prompt length by itself, is the common feature of the conditions with large statistical-score shifts. Appendix B.7 gives the prompt audit, generator strata, classifier floor analysis, and reproduction bounds.

4.6 Generator identity changes the strength of the shift

The pooled shift has distinct generator profiles (Figure 1). GLM's median standardized A–P1 contrast is 1.007, compared with 0.063 for MiniMax; some outputs also change direction across generators. The follow-ups likewise separate strong and weak responses, including the large B/C shifts for DeepSeek and GLM and the small shifts for Kimi-K2.5 in Experiment V. Reporting generator strata alongside the pooled average shows where the contextual response is strongest.

5 Related work and implications

Detection research develops both supervised representations and statistical criteria, including GLTR, DetectLLM, Fast-DetectGPT, LastDE, and Binoculars. Robustness benchmarks test how these approaches respond to changes in generation and editing. RAID covers diverse generators, decoding configurations, and adversarial perturbations (Dugan et al., 2024); Beemo studies expert editing of machine outputs (Artemova et al., 2025). Our intervention enters before generation: it changes the supplied context while preserving the requested topic.

The closest comparisons concern prompting and personalization. Koike et al. (2024) show that natural task constraints influence detection. Li & Wan (2025) study author attributes and detection

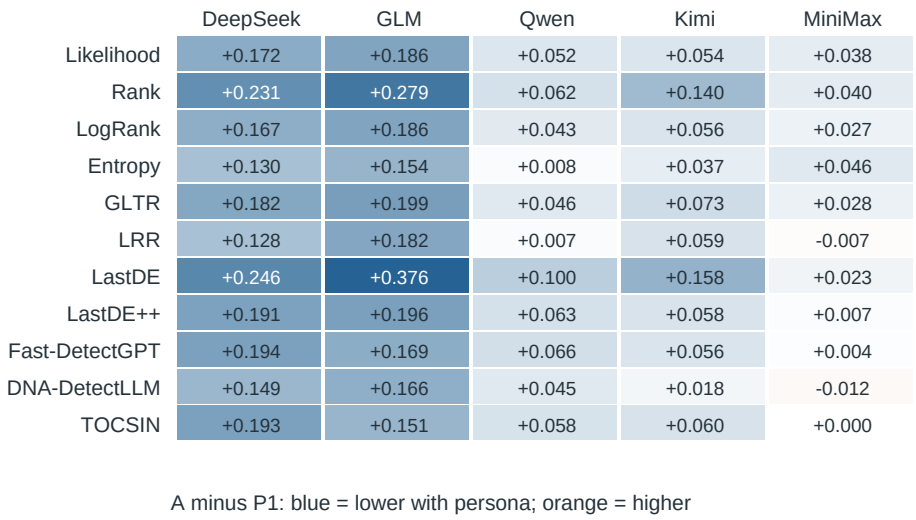

Figure 1: Experiment I: generator-specific A minus P1 contrasts for the 11-statistic subset. Each cell averages 150 paired topics. Blue positive values indicate lower persona scores; orange negative values indicate the opposite direction. The pooled effect has distinct generator-specific profiles.

bias. Gao et al. (2026) construct StyloBench using human literary and blog texts paired with generated imitations and analyze feature inversion. Our comparison places a detailed background, a brief role label, and a direct anti-AI-style instruction within the same topic–generator cells, then adds a detailed setting with no assigned writer. It establishes their relative score responses and generator dependence. Prompt-steerability research offers broader context for how instructions alter generated behavior (Miehling et al., 2025).

Implications for detector evaluation. Prompt composition is an evaluation variable. A one-sentence identity represents a different generation condition from a detailed writer history, and the non-biographical town context shows that an explicit writer role is unnecessary for a large statistical-score response. Detector evaluations should vary the amount and type of concrete context supplied to generators, distinguish these conditions from direct style directives, and report generator-specific score responses.

6 Conclusion

In the tested Chinese literary-writing setting, richly developed context produces much larger detector-score shifts than a brief writer label. Life-history facts carry the larger component response within a detailed persona, while a fictional town description without an assigned writer produces a similar response to those facts in a separate, internally paired experiment. Explicit writer identity is therefore unnecessary for the large statistical-score shift observed here. Detector evaluations should treat contextual detail as a prompt axis in its own right and report its effect across generators.

Reproducibility Statement

Section 3 specifies corpus construction, sampling settings, and inference units. Appendices A–D provide scoring configurations, complete principal contrasts, corpus overlap, component rules, resampling summaries, the construct-control audit, and feature definitions. Appendix E reproduces the recorded Chinese prompts. The accompanying code rebuilds the tables and vector figures from saved scores; `build_construct_validity.py` additionally re-derives every prompt template from the generation records and asserts the corpus properties quoted in the text. The analysis files retain per-topic estimands.

Ethics statement

The study uses generated passages and published literary excerpts, with no new human participants. The fictional personas are constructed descriptions; R uses public biographical information. Detector sensitivity has implications for reliable attribution and potential evasion. We present the findings to improve evaluation; individual authorship decisions require evidence beyond a detector score. Human-excerpt provenance is recorded, with redistribution governed by the source permissions.

Statement of AI use

Generative AI produced the synthetic passages studied here and the fixed fictional writer descriptions. AI assistants helped develop the paper's conceptual framing, refine hypotheses and experimental controls, implement and review analysis code, interpret results, search and summarize literature, prepare figures, translate research material, and draft and edit the manuscript. The authors reviewed the AI-assisted text, code, figures, and references, checked numerical claims against saved generation and scoring records and analysis outputs, and made the final decisions about the study design and conclusions. The authors take responsibility for the final content, including AI-assisted claims and artifacts.

References

Ekaterina Artemova, Jason S Lucas, Saranya Venkatraman, Jooyoung Lee, Sergei Tilga, Adaku Uchendu, and Vladislav Mikhailov. Beemo: Benchmark of expert-edited machine-generated outputs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6992–7018, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.357. URL https://aclanthology.org/2025.naacl-long.357/.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bpcgcr8E8Z.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.674. URL https://aclanthology.org/2024.acl-long.674/.

Lang Gao, Xuhui Li, Chenxi Wang, Mingzhe Li, Wei Liu, Zirui Song, Jinghui Zhang, Rui Yan, Preslav Nakov, and Xiuying Chen. When personalization tricks detectors: The feature-inversion trap in machine-generated text detection. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43143–43171, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1998. URL https://aclanthology.org/2026.acl-long.1998/.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In Marta R. Costa-jussà and Enrique Alfonseca (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019. URL https://aclanthology.org/P19-3019/.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17519–17537, 2024. URL https://proceedings.mlr.press/v235/hans24a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. How you prompt matters! Even task-oriented constraints in instructions affect LLM-generated text detection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14384–14395, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.841. URL https://aclanthology.org/2024.findings-emnlp.841/.

Jiatao Li and Xiaojun Wan. Who writes what: Unveiling the impact of author roles on ai-generated text detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://arxiv.org/abs/2502.12611.

Shixuan Ma and Quan Wang. Zero-shot detection of LLM-generated text using token cohesiveness. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17538–17553, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.971. URL https://aclanthology.org/2024.emnlp-main.971/.

Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M. Daly, Kush R. Varshney, Eitan Farchi, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, Miao Liu, and Prasanna Sattigeri. Evaluating the prompt steerability of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7874–7900, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.400. URL https://aclanthology.org/2025.naacl-long.400/.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.827. URL https://aclanthology.org/2023.findings-emnlp.827/.

Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. Training-free llm-generated text detection by mining token probability sequences. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.06072.

Xiaowei Zhu, Yubing Ren, Fang Fang, Qingfeng Tan, Shi Wang, and Yanan Cao. Dna-detectllm: Unveiling ai-generated text via a dna-inspired mutation-repair paradigm. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0287c393907259ae6a269fca5e3509cd-Abstract-Conference.html.

A Implementation and corpus provenance

A.1 Generation records and reuse

All recorded calls use DashScope. Experiments I–III use the exact identifiers `deepseek-v3`, `glm-4.5-air`, `Moonshot-Kimi-K2-Instruct`, `MiniMax-M2.1`, and `qwen3-32b`; Experiments II and III use the DeepSeek and Kimi identifiers from that set. Experiment IV uses `deepseek-v3.1`, `kimi-k2.5`, and `glm-4.5`, with reasoning disabled. Temperature, top-p, and output-token cap are 0.9, 0.9, and 900 for every generated record. The order seed controls request ordering and is not treated as a provider-enforced sampling seed.

There are 4,500 main-experiment texts. Experiment II contains 90 original P1 texts and 180 neutralized/removed texts. Experiment III contains 360 original texts and 120 P2/P3 texts. Experiment IV-A adds 360 N/B/T/P1 texts from the new three-generator cohort. IV-B contains 540 N/A/P1 records over two draws, of which 180 draw-1 N/P1 records are reused from IV-A and 360 are new. Experiment V contains 450 independently generated N/B/C/A/A' records. Literal text and design-key comparisons therefore give $4,500 + 180 + 120 + 360 + 360 + 450 = 5,970$ distinct generation records. Within each scoring corpus there is exactly one row per text and detector output, with no missing score and all normalized scores in $[0, 1]$. The score counts are 111,600, 6,480, 11,520, 8,640, 12,960, and 10,350 for Experiments I, II, III, IV-A, IV-B, and V, respectively. The first count includes 3,600 human-reference scores.

A.2 Scoring configurations

The two trained classifiers are `yuchuantian/AIGC_detector_zhv2` and `Hello-SimpleAI/chatgpt-detector-roberta-chinese`. Label 1 maps to AI and label 0 to human. Their default aggregation is a length-weighted mean over sliding windows of 510 tokens with stride 256; the tokenizer maximum is 512 tokens including special tokens. Four alternatives are recorded: beginning-only, ending-only, unweighted window mean, and window maximum. These are alternative outputs of the same classifier, not additional trained detectors.

MGT-Eval denotes the implementation codebase used to compute the 12 statistical outputs, not a dataset or an additional learned detector. The local manifests pin its recorded source revision and configurations for Experiment I and the follow-ups. The scoring/reference model is `uer/gpt2-chinese-cluecorpussmall`; the follow-up manifests specify a 512-token maximum for the recorded statistical methods. The main manifest does not supply the same complete token-limit metadata, so we do not infer that all main-experiment statistics were computed over entire passages. Classifier sliding-window results and statistical-method results retain their separate preprocessing definitions.

The adapter returns negative mean rank and negative mean log-rank. Hence larger raw Rank and LogRank outputs already correspond to the designated AI-oriented direction. Entropy is sign-reversed before ranking; the other recorded statistics use higher-is-AI orientation. Binoculars returns a signed implementation score and uses identical observer/performer identifiers in these runs. No sign is selected to maximize the observed persona contrast or reference AUROC in the present reanalysis. The manuscript uses the saved audited scores and does not rerun or recalibrate a detector.

The style proxy is a deterministic full-text score, not a trained authorship classifier. Let T be the total occurrence count over the 15 template strings listed in `src/detectors/heuristic_features.py`, U the count over its 10 summary-ending strings, σ the population standard deviation of sentence-segment lengths, and ℓ their mean. Define $m(\ell) = 0.8$ for $15 \le \ell \le 25$, $m(\ell) = 0.5$ for $10 \le \ell < 15$ or $25 < \ell \le 30$, and $m(\ell) = 0.2$ otherwise. Define $q = 0.7$ when none of the Chinese curly or ASCII single/double quote markers occurs and $q = 0.3$ otherwise. The reported version `handcrafted_features_v3_quote_fix` is

$$a_{\text{style}}(x) = 0.30\min(T/10, 1) + 0.25\min(U/5, 1) + 0.20\max(0, 1-\sigma/30) + 0.15m(\ell) + 0.10q. \quad (3)$$

The weights sum to one and the implementation clips the final value to $[0, 1]$ without corpus-level normalization. The released matrices `evidence/correlations_I.csv`, `correlations_II.csv`, and `correlations_III.csv` document that several statistical signals co-vary strongly; no independence assumption over outputs is used for a combined significance claim. Holm adjustment controls its specified family without requiring independent tests.

A.3 Two different ensemble constructions

For Experiment I, the saved ensemble uses supervised logistic stacking. Machine passages are partitioned by generator family; each human reference is deterministically assigned to one of those folds. Each held-out fold is scored using a logistic model fit to the other folds, with training-fold standardization, inverse-class-frequency weighting, and an L_2 penalty. The 15 base outputs are the two default classifiers, 12 statistics, and style proxy. Cross-fitting folds are defined by generator family, and the underlying statistical mid-ranks use the common corpus reference.

Experiments II–IV use a machine-only ensemble. For each held-out generator, each base output is transformed by the empirical CDF of the other generator or generators, then the 15 transformed values are averaged. The saved metadata call this `unsupervised_training_ecdf_rank_mean`. Its values measure relative score position. Both ensemble constructions are supplementary to the individual scoring outputs.

A.4 Human reference and interpretation

The 150 external excerpts are retrieved from modern Chinese literature. The source metadata record 71 excerpts by Yu Dafu, 36 by Lu Xun, 28 by Zhu Ziqing, 7 by Ye Zi, 6 by Xiao Hong, and 2 by Su Manshu. The excerpts provide an external literary reference, selected by topic retrieval. The exact fixed excerpts and row-level author, title, URL, matched-concept, match-score, coverage, and quality fields are released in `data/human/human_baseline_modern_literature_v2.csv`; downstream reproduction uses these rows rather than rerunning retrieval. Processing records describe contiguous excerpts, conversion of traditional to simplified characters, and cleaning of transcription or editorial markers without rewriting the prose.

AUROC compares the fixed-orientation machine scores with this reference. It is computed as

$$\widehat{\mathrm{AUC}}_d(c) = \frac{1}{n_c n_H} \sum_{i=1}^{n_c} \sum_{j=1}^{n_H} \left[\mathbb{I}(a_d(x_i) > a_d(h_j)) + \tfrac{1}{2}\mathbb{I}(a_d(x_i) = a_d(h_j)) \right]. \tag{4}$$

The human excerpts are shared across conditions. Table 6 reports descriptive AUROCs under the fixed score orientation; values below 0.5 indicate reversed ranking relative to this reference. In particular, LastDE and LRR are below 0.5 under N, and their persona score reductions are not evidence of becoming more human-like. No direction is reselected using this reference. Statistical tests in the main text concern paired machine-score contrasts. These historical excerpts are not matched contemporary human responses to the writing tasks; the AUROCs characterize separation from this particular reference, not deployment performance. The strong ensemble AUROC under P1 (0.981, versus 0.990 under N) also shows why score changes should be distinguished from loss of ranking performance.

Table 6: Experiment I AUROC by prompt condition. Each column compares 750 machine passages with the same 150 human excerpts; machine text is the positive class.

Output	N	S	A	P1	P1+A	R
AIGC-ZHv2	0.968	0.967	0.904	0.902	0.870	0.949
AIGC / beginning only	0.964	0.959	0.897	0.877	0.841	0.938
AIGC / ending only	0.966	0.963	0.899	0.888	0.852	0.942
AIGC / window max	0.964	0.962	0.903	0.901	0.878	0.944
AIGC / window mean	0.965	0.963	0.902	0.897	0.873	0.943
RoBERTa-ZH	0.913	0.910	0.879	0.843	0.850	0.864
RoBERTa / beginning only	0.910	0.904	0.862	0.818	0.813	0.849
RoBERTa / ending only	0.906	0.901	0.858	0.808	0.810	0.844
RoBERTa / window max	0.906	0.901	0.878	0.865	0.870	0.870
RoBERTa / window mean	0.906	0.901	0.880	0.855	0.862	0.866
Ensemble	0.990	0.988	0.983	0.981	0.977	0.979
Style proxy	0.839	0.828	0.858	0.750	0.804	0.768
Binoculars (same LM)	0.890	0.878	0.879	0.810	0.835	0.832
DNA-DetectLLM	0.899	0.890	0.876	0.829	0.843	0.849
Entropy	0.677	0.673	0.696	0.638	0.656	0.577

Output	N	S	A	P1	P1+A	R
Fast-DetectGPT	0.889	0.881	0.883	0.806	0.831	0.824
GLTR	0.689	0.675	0.717	0.612	0.659	0.580
LastDE	0.457	0.456	0.566	0.349	0.498	0.386
LastDE++	0.834	0.815	0.820	0.728	0.758	0.757
Likelihood	0.811	0.803	0.829	0.760	0.788	0.695
LogRank	0.727	0.715	0.749	0.662	0.697	0.614
LRR	0.429	0.418	0.429	0.332	0.374	0.359
Rank	0.672	0.649	0.740	0.585	0.673	0.557
TOCSIN	0.872	0.862	0.870	0.793	0.821	0.808

B Complete principal contrasts

B.1 Experiment I: persona shifts against both baselines

Tables 7 and 8 report all 24 outputs, averaging five generators within each of 150 topics. Positive values indicate lower scores under P1. Pointwise confidence intervals come from 5,000 topic bootstrap resamples with reanalysis seed 20260913. All p values are two-sided and Holm-adjusted across the 24 outputs separately for each baseline contrast. The tables express both contrasts as baseline minus P1; the evidence CSV files preserve their archived P1-minus-baseline signs.

Table 7: Full Experiment I N–P1 contrast.

Output	Difference	95% interval	d_z	Holm p
AIGC-ZHv2	0.2365	[0.2173, 0.2560]	1.933	< 0.001
AIGC / beginning only	0.2255	[0.2060, 0.2456]	1.800	< 0.001
AIGC / ending only	0.2396	[0.2190, 0.2602]	1.798	< 0.001
AIGC / window max	0.2802	[0.2550, 0.3062]	1.735	< 0.001
AIGC / window mean	0.2243	[0.2050, 0.2442]	1.812	< 0.001
RoBERTa-ZH	0.0003	[0.0001, 0.0005]	0.236	< 0.001
RoBERTa / beginning only	0.0001	[0.0001, 0.0001]	0.384	< 0.001
RoBERTa / ending only	0.0001	[0.0001, 0.0001]	0.366	< 0.001
RoBERTa / window max	0.0035	[0.0007, 0.0071]	0.167	< 0.001
RoBERTa / window mean	0.0012	[0.0003, 0.0025]	0.173	< 0.001
Ensemble	0.0777	[0.0642, 0.0918]	0.882	< 0.001
Style proxy	0.0374	[0.0307, 0.0442]	0.902	< 0.001
Binoculars (same LM)	0.1455	[0.1313, 0.1598]	1.623	< 0.001
DNA-DetectLLM	0.1402	[0.1259, 0.1545]	1.519	< 0.001
Entropy	0.0598	[0.0472, 0.0723]	0.768	< 0.001
Fast-DetectGPT	0.1446	[0.1303, 0.1588]	1.623	< 0.001
GLTR	0.0995	[0.0885, 0.1112]	1.396	< 0.001
LastDE	0.0835	[0.0627, 0.1047]	0.641	< 0.001
LastDE++	0.1534	[0.1385, 0.1683]	1.661	< 0.001
Likelihood	0.1125	[0.1018, 0.1237]	1.621	< 0.001
LogRank	0.1035	[0.0924, 0.1149]	1.482	< 0.001
LRR	0.0650	[0.0512, 0.0787]	0.756	< 0.001
Rank	0.0970	[0.0834, 0.1110]	1.083	< 0.001
TOCSIN	0.1356	[0.1214, 0.1503]	1.481	< 0.001

Table 8: Full Experiment I A–P1 contrast.

Output	Difference	95% interval	d_z	Holm p
AIGC-ZHv2	0.0621	[0.0473, 0.0773]	0.664	< 0.001
AIGC / beginning only	0.0674	[0.0523, 0.0817]	0.729	< 0.001
AIGC / ending only	0.0643	[0.0496, 0.0791]	0.686	< 0.001
AIGC / window max	0.0568	[0.0358, 0.0776]	0.436	< 0.001
AIGC / window mean	0.0555	[0.0395, 0.0712]	0.557	< 0.001
RoBERTa-ZH	0.0006	[0.0002, 0.0012]	0.178	< 0.001

Output	Difference	95% interval	d_z	Holm p
RoBERTa / beginning only	0.0005	[0.0001, 0.0011]	0.129	< 0.001
RoBERTa / ending only	0.0005	[0.0001, 0.0011]	0.128	< 0.001
RoBERTa / window max	0.0033	[0.0005, 0.0071]	0.155	< 0.001
RoBERTa / window mean	0.0016	[0.0004, 0.0033]	0.166	< 0.001
Ensemble	0.0287	[0.0118, 0.0446]	0.284	< 0.001
Style proxy	0.0373	[0.0319, 0.0429]	1.056	< 0.001
Binoculars (same LM)	0.0963	[0.0812, 0.1117]	1.003	< 0.001
DNA-DetectLLM	0.0732	[0.0573, 0.0885]	0.742	< 0.001
Entropy	0.0751	[0.0603, 0.0896]	0.819	< 0.001
Fast-DetectGPT	0.0980	[0.0834, 0.1125]	1.051	< 0.001
GLTR	0.1054	[0.0923, 0.1183]	1.312	< 0.001
LastDE	0.1805	[0.1612, 0.2005]	1.447	< 0.001
LastDE++	0.1029	[0.0878, 0.1183]	1.069	< 0.001
Likelihood	0.1003	[0.0883, 0.1129]	1.284	< 0.001
LogRank	0.0958	[0.0829, 0.1081]	1.223	< 0.001
LRR	0.0740	[0.0591, 0.0897]	0.768	< 0.001
Rank	0.1503	[0.1368, 0.1637]	1.787	< 0.001
TOCSIN	0.0925	[0.0771, 0.1076]	0.961	< 0.001

B.2 Experiment I: brief writer-label controls

These additional comparisons use the saved main-study scores, average the five generators within each of 150 topics, and apply two-sided Wilcoxon tests. Each contrast retains a 24-output Holm family, including the eight alternative classifier aggregations. Table 9 shows the 16 default outputs once. Positive N–S indicates lower scores with the brief label; positive S–P1 indicates lower scores with the biography. Pointwise intervals in `evidence/strategy_controls_I.csv` use 5,000 topic bootstrap resamples and seed 20260917; the same file includes every aggregation and generator stratum.

Table 9: Brief label versus neutral prompting, and brief label versus detailed background. The outputs share models and score constructions.

	N–S			S–P1		
Output	Difference	d_z	Holm p	Difference	d_z	Holm p
AIGC-ZHv2	0.0130	0.112	1.000	0.2235	1.792	< 0.001
RoBERTa-ZH	0.0001	0.098	0.664	0.0001	0.293	< 0.001
Style proxy	0.0067	0.177	0.562	0.0307	0.782	< 0.001
Ensemble	0.0156	0.184	0.438	0.0621	0.638	< 0.001
Likelihood	0.0111	0.144	1.000	0.1014	1.325	< 0.001
Rank	0.0178	0.218	0.257	0.0792	0.913	< 0.001
LogRank	0.0099	0.131	1.000	0.0936	1.232	< 0.001
Entropy	0.0011	0.016	1.000	0.0586	0.701	< 0.001
LRR	0.0054	0.061	1.000	0.0597	0.672	< 0.001
GLTR	0.0096	0.122	1.000	0.0899	1.145	< 0.001
LastDE	0.0030	0.025	1.000	0.0805	0.619	< 0.001
LastDE++	0.0185	0.195	0.602	0.1349	1.353	< 0.001
Fast-DetectGPT	0.0174	0.191	0.602	0.1272	1.417	< 0.001
DNA-DetectLLM	0.0185	0.202	0.602	0.1217	1.304	< 0.001
Binoculars (same LM)	0.0207	0.231	0.336	0.1248	1.366	< 0.001
TOCSIN	0.0167	0.179	0.664	0.1189	1.284	< 0.001

B.3 Experiment III: individual persona comparisons

Table 10: Experiment III: shared direction and persona variation across the 14-signal panel. Positive contrasts mean the first condition receives higher scores. $\overline{P}$ averages the three fixed personas within each topic and generator. Counts describe dependent outputs; Holm correction is applied separately for each contrast.

Contrast	Positive outputs	Holm significant	Median d_z
N–P1	14/14	12/14	0.776
N–P2	14/14	14/14	1.426
N–P3	14/14	13/14	1.206
N–$\overline{P}$	14/14	13/14	1.382
A–$\overline{P}$	14/14	14/14	1.295
S–$\overline{P}$	14/14	10/14	0.815
R–$\overline{P}$	9/14	1/14	0.244
P1–P2	14/14	13/14	0.842
P1–P3	11/14	0/14	0.380
P3–P2	14/14	7/14	0.486

Table 11 gives the five individual-persona contrasts used in the main analysis.

Table 11: Topic-averaged differences in Experiment III. * denotes Holm-adjusted $p < 0.05$ within the 14-output family for that column's contrast. The three neutral contrasts and the two P2 comparisons retain distinct inferential families.

Output	N–P1	N–P2	N–P3	P1–P2	P3–P2
Likelihood	0.157*	0.365*	0.256*	0.208*	0.109*
Rank	0.147*	0.276*	0.145*	0.129*	0.131*
LogRank	0.148*	0.344*	0.235*	0.196*	0.109*
Entropy	0.090*	0.209*	0.138*	0.119*	0.071
GLTR	0.146*	0.332*	0.220*	0.186*	0.112*
LRR	0.087	0.215*	0.142*	0.128*	0.072
LastDE	0.051	0.124*	0.033	0.073	0.092
LastDE++	0.216*	0.412*	0.309*	0.196*	0.103*
Fast-DetectGPT	0.180*	0.355*	0.281*	0.175*	0.074
DNA-DetectLLM	0.168*	0.348*	0.273*	0.179*	0.075
TOCSIN	0.160*	0.322*	0.240*	0.163*	0.082
Binoculars (same LM)	0.181*	0.346*	0.271*	0.164*	0.074
Style proxy	0.059*	0.094*	0.065*	0.035*	0.029*
Ensemble	0.161*	0.313*	0.234*	0.152*	0.079*

B.4 Experiment II: component intervention

The component contrasts average two generators within each of 45 topics. PN replaces P1's explicit temperament and observational fragments with character-matched neutral facts; PR removes those fragments. Both retain the surrounding biography. P1–PN is positive on all 14 signals and P1–PR on 13 (Table 12); no individual comparison survives correction. The strongest P1–PR result is LogRank, with raw $p = 0.00445$ and corrected $p = 0.0623$. These local edits motivate the cleaner B/T partition in Experiment IV.

Table 12: Experiment II: explicit personality-block interventions on 45 paired topics. Positive differences indicate higher scores under the first condition.

Contrast	Positive	Median d_z	Holm significant
P1–PN	14/14	0.267	0/14
P1–PR	13/14	0.346	0/14
PN–PR	13/14	0.107	0/14

Table 13: All 14 component signals. The two contrasts are corrected separately.

	P1–PN			P1–PR	
Output	Difference	d_z	Holm p	Difference	Holm p
Likelihood	0.048	0.288	0.671	0.089	0.135
Rank	0.018	0.085	1.000	0.096	0.135
LogRank	0.062	0.355	0.428	0.098	0.062
Entropy	0.026	0.138	1.000	0.051	0.237
GLTR	0.069	0.361	0.230	0.090	0.135
LRR	0.073	0.312	0.671	0.101	0.135
LastDE	0.043	0.185	1.000	0.070	0.344
LastDE++	0.058	0.249	0.896	0.086	0.227
Fast-DetectGPT	0.061	0.263	0.896	0.083	0.344
DNA-DetectLLM	0.056	0.241	0.896	0.080	0.344
TOCSIN	0.063	0.278	0.767	0.076	0.344
Binoculars (same LM)	0.063	0.271	0.896	0.079	0.344
Style proxy	0.001	0.014	1.000	-0.001	0.902
Ensemble	0.044	0.337	0.428	0.062	0.111

B.5 Experiment IV-A: deterministic B/T decomposition

B and T are literal subsets of the P1 prefix. B retains age, family, education, employment, travel, relationships, lived settings, and reading history. It removes or minimally repairs seven fragments that specify observational habits, interpretations, thematic interests, temperament, or narrative aesthetics. T concatenates the removed cue-bearing renderings without adding a life event. The shared task suffix is unchanged. The seven audited operations are: separating travel from recording accents, gestures, and silence; bereavement objects from interpreting silence; a relationship event from admitting weakness; assigning the full anti-slogan sentence to T; separating lived settings from thematic interests; reading history from the anti-display stance; and assigning the final temperament/observation paragraph to T. The released `src/persona_bt_partition.py` stores each literal source, B rendering, T rendering, and rationale.

Table 14 reports the four estimands emphasized in the main text for every output in the 14-signal family. Asterisks mark Holm-adjusted $p < 0.05$ within the row's estimand family. Positive values denote score reductions. The interaction is $B+T-N-P1$; its negative values indicate sub-additivity on the score scale, not harmful interaction in generated-text quality.

Table 14: Experiment IV-A, three-generator average over 30 topics.

Output	Biography main	Style main	B–P1	B×T interaction
Likelihood	0.229*	0.046	-0.020	-0.133*
Rank	0.198*	0.014	-0.045	-0.118
LogRank	0.218*	0.042	-0.024	-0.133
Entropy	0.163*	0.018	-0.055	-0.147*
GLTR	0.211*	0.034	-0.021	-0.110
LRR	0.145*	0.019	-0.027	-0.093
LastDE	0.128*	-0.011	-0.027	-0.032
LastDE++	0.235*	0.061	-0.001	-0.124
Fast-DetectGPT	0.226*	0.067*	0.008	-0.118
DNA-DetectLLM	0.226*	0.076*	0.015	-0.122
Binoculars (same-LM)	0.230*	0.067	0.008	-0.116
TOCSIN	0.227*	0.061	0.007	-0.108
Style proxy	0.036*	-0.002	-0.006	-0.008
Rank ensemble	0.185*	0.043	-0.015	-0.118*

The ten classifier/window outputs are retained as a separate supplementary analysis family rather than expanding the 14-test family. Across all 24 outputs, the life-fact main effect is positive on 24 and corrected-significant on 22; the narrative/style main effect is positive on 22 and corrected-significant on 10. B–P1 is positive on eight and corrected-significant on none.

B.6 Experiment IV-B: independent stochastic resampling

Table 15 summarizes the two draws and their within-cell average. Table 16 gives every draw-specific contrast in the 14-signal family. Asterisks denote Holm-adjusted $p < 0.05$ within each contrast and draw. The two-draw mean first averages scores within topic–generator–condition cells, then forms topic-level contrasts; it does not double the topic sample size.

Table 15: Experiment IV-B aggregate resampling results. Positive counts denote lower P1 scores.

Contrast	Positive	Holm sig.	Median reduction	Median d_z
N–P1, draw 1	14/14	14/14	0.277	1.979
N–P1, draw 2	14/14	13/14	0.201	1.411
N–P1, two-draw mean	14/14	14/14	0.239	2.206
A–P1, draw 1	14/14	14/14	0.291	1.709
A–P1, draw 2	14/14	14/14	0.223	1.627
A–P1, two-draw mean	14/14	14/14	0.253	2.012

Table 16: Experiment IV-B draw-specific reductions.

	N–P1		A–P1	
Output	Draw 1	Draw 2	Draw 1	Draw 2
Likelihood	0.300*	0.222*	0.325*	0.280*
Rank	0.225*	0.154*	0.333*	0.323*
LogRank	0.287*	0.215*	0.326*	0.287*
Entropy	0.192*	0.156*	0.241*	0.266*
GLTR	0.267*	0.188*	0.332*	0.285*
LRR	0.182*	0.149*	0.279*	0.265*
LastDE	0.116*	0.038	0.276*	0.233*
LastDE++	0.311*	0.212*	0.288*	0.200*
Fast-DetectGPT	0.310*	0.223*	0.293*	0.200*
DNA-DetectLLM	0.316*	0.236*	0.269*	0.181*
Binoculars (same-LM)	0.312*	0.229*	0.296*	0.201*
TOCSIN	0.305*	0.223*	0.300*	0.204*
Style proxy	0.034*	0.043*	0.039*	0.042*
Rank ensemble	0.238*	0.189*	0.248*	0.214*

B.7 Experiment V: construct-validity control

`build_construct_validity.py` re-derives every quantity in this appendix from the saved generation and scoring records and raises if any assertion fails. Batch `construct_validity_control_20260918` contains 450 texts, 90 per condition across N, B, C, A, and A′, over the same 30 topics and the same three generator identifiers as Experiment IV-A. All 450 calls ran through DashScope against `deepseek-v3.1`, `glm-4.5`, and `kimi-k2.5` with temperature 0.9, top-p 0.9, a 900-token output cap, an empty system prompt, and no `api_seed` recorded; generation ran from 2026-09-18T13:11:00Z to 2026-09-18T15:27:09Z. Scoring produced 10,350 rows—23 outputs for each of the 450 texts—with no duplicate `text_id`–`detector_name` pair and no missing `ai_score`. The script recovers the prompt template for each condition by replacing the literal topic string in every saved `user_prompt`, asserts that exactly one template per condition results, and asserts that the recovered N and B templates are byte-identical to the N and B templates recovered from the Experiment IV-A records. The two experiments therefore share prompt wording for those conditions and share no generated text.

Six of the 450 generations ended with `finish_reason=length` rather than `stop`: four under C and two under B, all from Kimi-K2.5 and truncated at the 900-token cap. Mean rendered prompt lengths are 123 (N), 594 (C), 667 (B), 860 (A′), and 902 (A) characters. Mean output lengths are 568.9 (N), 655.5 (C), 653.8 (B), 525.8 (A′), and 515.6 (A); medians are 572.0, 609.0, 619.5, 539.5, and 551.5.

The A′ variant as recorded. A′ was authored by removing A's factuality bullet. The recorded template also contains one unrelated character substitution: 衔接词 is written as 衍接词. The deleted 42-character bullet reads 避免虚假：除非我明确允许虚构，否则不要生成不存在的、不真实的、未经证实的内容。 The exact template diff is stored in `evidence/corpus_audit_V.json`. The same audit confirms that V's A template matches Experiment I's and that V's N and B templates match Experiment IV-A's byte for byte.

Rank reference. The 12 statistical outputs store an ECDF midrank percentile computed over all scored rows for that output within this batch. The script parses `normalization_n` out of each saved `raw_output` record and asserts that it equals 450 for every statistical row. Experiment V's rank values therefore sit on their own 450-text reference and are not comparable with the 360-text reference of IV-A or the 540-text reference of IV-B. No value in this appendix is differenced against a value from another experiment.

Equivalence margins that this design could support. For each of the 12 statistical outputs we ask whether the entire B–C interval lies inside $\pm m \cdot |\widehat{\Delta}_d(N, B)|$ for that same output, which is the two-one-sided-test criterion for equivalence at margin m. The counts passing are 0 at $m = 0.20$, 3 at $m = 0.25$, 7 at $m = 0.33$, and 11 at $m = 0.50$. The median B–C interval halfwidth is 0.059 against a median $|N - B|$ of 0.257, so the design resolves roughly a quarter of the effect it is trying to bound. A ±20% equivalence claim is out of reach on every output, and we make none.

Trained classifier floors. Under the default aggregation, AIGC-ZHv2's median score is 0.198 for N but 0.0027 for B, where 95.6% of B texts fall below 0.05; C sits at 0.0112 with 71.1% below, A′ at 0.0074 with 73.3%, and A at 0.0062 with 81.1%. Because B is at the detector's floor, the ordering B, A, C on this scale reflects saturation rather than graded sensitivity, and the corrected-significant AIGC B–C contrast of 0.0970 cannot be read as C lying meaningfully closer to human text. RoBERTa-ZH returns a median of at most 0.0001 in all five conditions with 100% of texts below 0.05, so it carries no usable information in this cohort and no claim about it rises above the single-output level. No human reference was scored in this batch, so Experiment V reports no AUROC; the AUROC values elsewhere in this paper belong to Experiment I's 150-excerpt historical reference and do not transfer.

Generator strata. Table 17 gives the median reduction over the 12 statistical outputs and the count of corrected-significant outputs in each stratum. Holm correction is applied within each contrast and stratum separately over the 13-signal analysis family. The pooled stratum averages the three generators within each topic before forming contrasts, so it weights the near-absent Kimi response equally with the two large ones.

Table 17: Experiment V median reductions over the 12 statistical outputs, and counts of Holm-significant outputs out of 12, by generator stratum. Positive values indicate lower scores than the second-named condition.

	Median reduction				Holm significant			
Contrast	Pooled	DeepSeek	GLM	Kimi	Pooled	DeepSeek	GLM	Kimi
N–B	0.257	0.358	0.333	0.041	12/12	12/12	12/12	0/12
N–C	0.264	0.438	0.259	0.081	12/12	12/12	12/12	0/12
N–A	-0.001	0.077	0.056	-0.094	0/12	0/12	0/12	6/12
N–A′	-0.083	-0.065	0.052	-0.201	6/12	0/12	0/12	7/12
A′–A	-0.070	-0.145	-0.008	-0.049	2/12	4/12	0/12	1/12
B–C	-0.012	-0.076	0.083	-0.041	0/12	1/12	0/12	0/12

Tables 18–23 report every output in all six contrasts of the pooled stratum, grouped into the 12 statistical outputs, the handcrafted style proxy, and the ten supplementary classifier and window aggregations. The 13-signal analysis family comprises the statistical outputs and the style proxy; the classifier rows form a separate supplementary family, as in Experiments IV-A and IV-B.

Table 18: Experiment V, N–B contrast.

Output	Difference	95% interval	d_z	Holm p
Statistical outputs				
Likelihood	0.2617	[0.2103, 0.3118]	1.799	< 0.001
Rank	0.2278	[0.1800, 0.2772]	1.642	< 0.001
LogRank	0.2542	[0.2012, 0.3049]	1.735	< 0.001
Entropy	0.1859	[0.1322, 0.2376]	1.238	< 0.001
GLTR	0.2342	[0.1788, 0.2873]	1.532	< 0.001
LRR	0.1995	[0.1378, 0.2573]	1.169	< 0.001
LastDE	0.1612	[0.1041, 0.2193]	1.006	< 0.001
LastDE++	0.2796	[0.2274, 0.3311]	1.880	< 0.001
Fast-DetectGPT	0.2635	[0.2139, 0.3117]	1.871	< 0.001
DNA-DetectLLM	0.2597	[0.2048, 0.3123]	1.687	< 0.001
Binoculars (same-LM)	0.2717	[0.2185, 0.3225]	1.812	< 0.001
TOCSIN	0.2632	[0.2132, 0.3121]	1.861	< 0.001
Auxiliary construction				
Style proxy	0.0631	[0.0417, 0.0857]	1.021	< 0.001
Trained classifiers				
AIGC-ZHv2	0.3792	[0.2990, 0.4575]	1.681	< 0.001
AIGC / beginning	0.3641	[0.2835, 0.4440]	1.610	< 0.001
AIGC / ending	0.4107	[0.3204, 0.4974]	1.664	< 0.001
AIGC / window max	0.4764	[0.3701, 0.5785]	1.621	< 0.001
AIGC / window mean	0.3497	[0.2740, 0.4227]	1.645	< 0.001
RoBERTa-ZH	0.0001	[0.0001, 0.0002]	0.683	< 0.001
RoBERTa / beginning	0.0001	[0.0000, 0.0001]	0.472	< 0.001
RoBERTa / ending	0.0001	[0.0000, 0.0001]	0.479	< 0.001
RoBERTa / window max	0.0027	[0.0002, 0.0073]	0.216	< 0.001
RoBERTa / window mean	0.0010	[0.0001, 0.0025]	0.233	< 0.001

Table 19: Experiment V, N–C contrast.

Output	Difference	95% interval	d_z	Holm p
Statistical outputs				
Likelihood	0.2647	[0.2008, 0.3326]	1.408	< 0.001
Rank	0.2534	[0.1940, 0.3158]	1.449	< 0.001
LogRank	0.2632	[0.1986, 0.3315]	1.390	< 0.001
Entropy	0.2041	[0.1385, 0.2702]	1.093	< 0.001
GLTR	0.2539	[0.1880, 0.3219]	1.357	< 0.001
LRR	0.2405	[0.1719, 0.3076]	1.251	< 0.001

Output	Difference	95% interval	d_z	Holm p
LastDE	0.2225	[0.1647, 0.2808]	1.369	< 0.001
LastDE++	0.2798	[0.2222, 0.3412]	1.615	< 0.001
Fast-DetectGPT	0.2743	[0.2192, 0.3302]	1.713	< 0.001
DNA-DetectLLM	0.2727	[0.2126, 0.3337]	1.560	< 0.001
Binoculars (same-LM)	0.2764	[0.2193, 0.3342]	1.677	< 0.001
TOCSIN	0.2735	[0.2167, 0.3312]	1.672	< 0.001
Auxiliary construction				
Style proxy	0.0773	[0.0586, 0.0962]	1.441	< 0.001
Trained classifiers				
AIGC-ZHv2	0.2822	[0.1913, 0.3698]	1.154	< 0.001
AIGC / beginning	0.2897	[0.2055, 0.3740]	1.242	< 0.001
AIGC / ending	0.3195	[0.2249, 0.4089]	1.246	< 0.001
AIGC / window max	0.3064	[0.1767, 0.4341]	0.862	0.001
AIGC / window mean	0.2336	[0.1401, 0.3243]	0.910	< 0.001
RoBERTa-ZH	-0.0003	[-0.0007, 0.0000]	-0.263	0.343
RoBERTa / beginning	0.0000	[-0.0001, 0.0001]	0.001	0.003
RoBERTa / ending	0.0000	[-0.0002, 0.0001]	-0.001	0.006
RoBERTa / window max	-0.0096	[-0.0310, 0.0042]	-0.178	0.579
RoBERTa / window mean	-0.0032	[-0.0103, 0.0014]	-0.179	0.579

Table 20: Experiment V, N–A contrast.

Output	Difference	95% interval	d_z	Holm p
Statistical outputs				
Likelihood	0.0022	[-0.0483, 0.0511]	0.016	1.000
Rank	-0.0839	[-0.1377, -0.0283]	-0.540	0.080
LogRank	-0.0048	[-0.0561, 0.0448]	-0.035	1.000
Entropy	-0.0392	[-0.1007, 0.0227]	-0.226	0.882
GLTR	-0.0277	[-0.0803, 0.0243]	-0.186	1.000
LRR	-0.0149	[-0.0736, 0.0404]	-0.091	1.000
LastDE	-0.0932	[-0.1549, -0.0302]	-0.517	0.109
LastDE++	0.0555	[-0.0010, 0.1162]	0.338	0.734
Fast-DetectGPT	0.0650	[0.0136, 0.1175]	0.446	0.404
DNA-DetectLLM	0.0909	[0.0355, 0.1474]	0.582	0.065
Binoculars (same-LM)	0.0679	[0.0131, 0.1239]	0.437	0.384
TOCSIN	0.0587	[0.0058, 0.1128]	0.393	0.513
Auxiliary construction				
Style proxy	0.0087	[-0.0136, 0.0315]	0.138	1.000
Trained classifiers				
AIGC-ZHv2	0.3341	[0.2534, 0.4136]	1.496	< 0.001
AIGC / beginning	0.3270	[0.2408, 0.4081]	1.402	< 0.001
AIGC / ending	0.3656	[0.2711, 0.4554]	1.401	< 0.001
AIGC / window max	0.4146	[0.3180, 0.5083]	1.567	< 0.001
AIGC / window mean	0.3052	[0.2311, 0.3768]	1.491	< 0.001
RoBERTa-ZH	0.0001	[0.0000, 0.0001]	0.242	0.230
RoBERTa / beginning	0.0000	[-0.0001, 0.0001]	0.060	0.037
RoBERTa / ending	0.0000	[-0.0001, 0.0001]	0.076	0.019
RoBERTa / window max	0.0016	[-0.0015, 0.0069]	0.127	1.000
RoBERTa / window mean	0.0006	[-0.0005, 0.0023]	0.135	1.000

Table 21: Experiment V, N–A′ contrast.

Output	Difference	95% interval	d_z	Holm p
Statistical outputs				
Likelihood	-0.0765	[-0.1267, -0.0254]	-0.531	0.019
Rank	-0.1574	[-0.2108, -0.1047]	-1.057	< 0.001
LogRank	-0.0901	[-0.1415, -0.0371]	-0.605	0.015
Entropy	-0.1205	[-0.1764, -0.0622]	-0.732	0.003

Output	Difference	95% interval	d_z	Holm p
GLTR	-0.1132	[-0.1645, -0.0632]	-0.781	0.002
LRR	-0.1168	[-0.1771, -0.0590]	-0.684	0.006
LastDE	-0.0904	[-0.1587, -0.0217]	-0.467	0.096
LastDE++	0.0173	[-0.0423, 0.0729]	0.108	1.000
Fast-DetectGPT	0.0049	[-0.0494, 0.0561]	0.033	1.000
DNA-DetectLLM	0.0281	[-0.0285, 0.0803]	0.182	0.948
Binoculars (same-LM)	0.0080	[-0.0476, 0.0598]	0.052	1.000
TOCSIN	-0.0080	[-0.0639, 0.0433]	-0.053	1.000
Auxiliary construction				
Style proxy	-0.0089	[-0.0293, 0.0100]	-0.163	1.000
Trained classifiers				
AIGC-ZHv2	0.3049	[0.2223, 0.3835]	1.334	< 0.001
AIGC / beginning	0.2922	[0.2079, 0.3759]	1.205	< 0.001
AIGC / ending	0.3410	[0.2534, 0.4259]	1.393	< 0.001
AIGC / window max	0.3848	[0.2868, 0.4806]	1.392	< 0.001
AIGC / window mean	0.2813	[0.2053, 0.3560]	1.328	< 0.001
RoBERTa-ZH	-0.0003	[-0.0009, 0.0001]	-0.202	0.532
RoBERTa / beginning	-0.0001	[-0.0004, 0.0001]	-0.107	0.213
RoBERTa / ending	-0.0001	[-0.0004, 0.0001]	-0.107	0.293
RoBERTa / window max	-0.0091	[-0.0330, 0.0061]	-0.149	0.809
RoBERTa / window mean	-0.0031	[-0.0110, 0.0019]	-0.152	0.809

Table 22: Experiment V, A′–A contrast.

Output	Difference	95% interval	d_z	Holm p
Statistical outputs				
Likelihood	-0.0787	[-0.1278, -0.0302]	-0.564	0.059
Rank	-0.0736	[-0.1231, -0.0229]	-0.523	0.089
LogRank	-0.0852	[-0.1334, -0.0381]	-0.630	0.039
Entropy	-0.0813	[-0.1403, -0.0262]	-0.507	0.157
GLTR	-0.0856	[-0.1361, -0.0356]	-0.591	0.066
LRR	-0.1019	[-0.1571, -0.0507]	-0.675	0.009
LastDE	0.0028	[-0.0520, 0.0523]	0.019	0.954
LastDE++	-0.0381	[-0.0894, 0.0101]	-0.270	0.954
Fast-DetectGPT	-0.0601	[-0.1169, -0.0067]	-0.384	0.552
DNA-DetectLLM	-0.0628	[-0.1187, -0.0102]	-0.401	0.552
Binoculars (same-LM)	-0.0599	[-0.1160, -0.0077]	-0.383	0.552
TOCSIN	-0.0667	[-0.1240, -0.0104]	-0.406	0.513
Auxiliary construction				
Style proxy	-0.0176	[-0.0371, 0.0018]	-0.313	0.552
Trained classifiers				
AIGC-ZHv2	-0.0292	[-0.0761, 0.0136]	-0.231	1.000
AIGC / beginning	-0.0348	[-0.0842, 0.0101]	-0.260	0.636
AIGC / ending	-0.0247	[-0.0746, 0.0235]	-0.180	1.000
AIGC / window max	-0.0298	[-0.0886, 0.0249]	-0.186	1.000
AIGC / window mean	-0.0239	[-0.0640, 0.0134]	-0.222	1.000
RoBERTa-ZH	-0.0003	[-0.0009, 0.0000]	-0.263	1.000
RoBERTa / beginning	-0.0001	[-0.0004, 0.0001]	-0.122	1.000
RoBERTa / ending	-0.0001	[-0.0004, 0.0001]	-0.126	1.000
RoBERTa / window max	-0.0107	[-0.0330, 0.0010]	-0.181	1.000
RoBERTa / window mean	-0.0037	[-0.0111, 0.0003]	-0.185	1.000

Table 23: Experiment V, B–C contrast.

Output	Difference	95% interval	d_z	Holm p
Statistical outputs				
Likelihood	-0.0030	[-0.0631, 0.0584]	-0.017	1.000
Rank	-0.0257	[-0.0845, 0.0328]	-0.159	1.000

Output	Difference	95% interval	d_z	Holm p
LogRank	-0.0090	[-0.0671, 0.0502]	-0.054	1.000
Entropy	-0.0182	[-0.0715, 0.0384]	-0.116	1.000
GLTR	-0.0196	[-0.0806, 0.0372]	-0.119	1.000
LRR	-0.0411	[-0.0980, 0.0160]	-0.251	1.000
LastDE	-0.0613	[-0.0991, -0.0231]	-0.568	0.099
LastDE++	-0.0001	[-0.0586, 0.0531]	-0.001	1.000
Fast-DetectGPT	-0.0108	[-0.0710, 0.0480]	-0.064	1.000
DNA-DetectLLM	-0.0130	[-0.0753, 0.0472]	-0.075	1.000
Binoculars (same-LM)	-0.0047	[-0.0650, 0.0534]	-0.028	1.000
TOCSIN	-0.0103	[-0.0706, 0.0496]	-0.061	1.000
Auxiliary construction				
Style proxy	-0.0142	[-0.0266, -0.0007]	-0.381	0.297
Trained classifiers				
AIGC-ZHv2	0.0970	[0.0531, 0.1483]	0.699	< 0.001
AIGC / beginning	0.0744	[0.0373, 0.1184]	0.628	< 0.001
AIGC / ending	0.0912	[0.0484, 0.1396]	0.682	< 0.001
AIGC / window max	0.1700	[0.0883, 0.2550]	0.714	< 0.001
AIGC / window mean	0.1161	[0.0644, 0.1752]	0.734	< 0.001
RoBERTa-ZH	0.0004	[0.0001, 0.0008]	0.387	0.066
RoBERTa / beginning	0.0001	[0.0000, 0.0002]	0.223	0.066
RoBERTa / ending	0.0001	[0.0000, 0.0002]	0.226	0.020
RoBERTa / window max	0.0123	[0.0007, 0.0322]	0.237	0.245
RoBERTa / window mean	0.0042	[0.0003, 0.0108]	0.241	0.153

C Additional persona and prose diagnostics

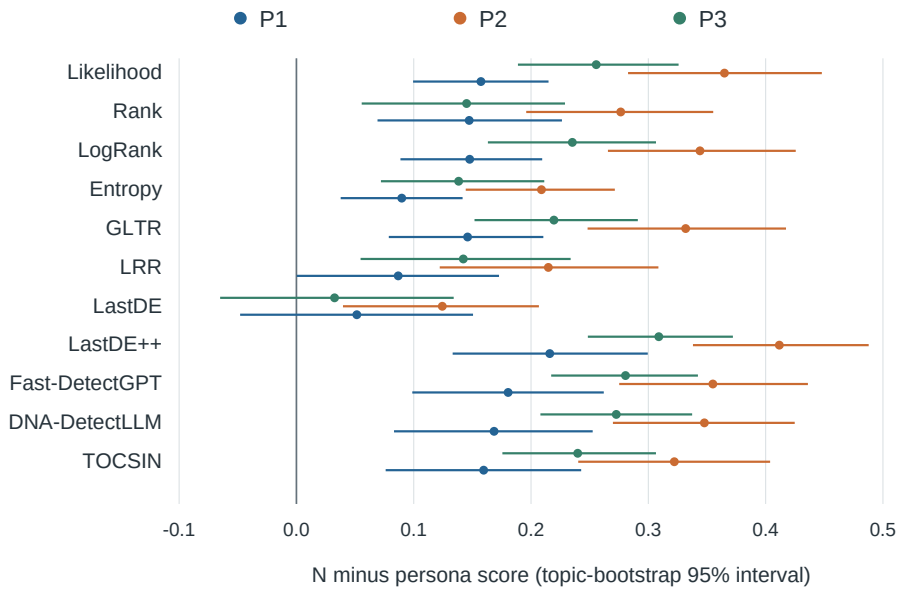

Figure 2: Experiment III: shared direction and differing magnitudes across three personas. Points are neutral-minus-persona contrasts, averaging DeepSeek and Kimi within 30 topics; bars are pointwise 95% topic-bootstrap intervals. All 11 non-auxiliary statistical outputs are shown. Positive values indicate lower persona scores.

P1 and A differ by 16 characters; P1 and P2 differ by ten. Large differences between these nearly equal-sized prompts make a simple "more characters produce lower scores" account inadequate. These complete-strategy comparisons alone do not isolate biographical semantics from wording, narrative cues, or instruction type; Experiment IV presents fixed component bundles, and Experiment V then tests the biographical reading of the life-fact bundle against a non-biographical context

Table 24: Matched excerpts from DeepSeek-V3 on the first numbered topic, "Remembering the light in the childhood kitchen on a rainy night." Chinese excerpts are contiguous and verbatim; English translations are for illustration and were not scored. Within this fixed topic, excerpts were chosen to show narrative detail, not by detector score or a quality rating.

Condition	Generated excerpt and English translation
S	母亲站在光晕中央，围裙上沾着面粉，锅铲与铁锅碰撞出清脆的声响。 Mother stood in the pool of light, flour on her apron; the spatula rang clearly against the iron wok.
A	祖母的锅铲碰着铁锅，铿铿地响，和雨声混在一起。她炒菜很少说话，偶尔叫我递盐罐子，手背上有块烫伤的疤。 Grandmother's spatula clanged against the iron wok, mingling with the rain. She rarely spoke while cooking, sometimes asking me to pass the salt jar; a burn scar marked the back of her hand.
P1	有一回父亲罕见地在家，他蹲在厨房角落修煤炉，工装裤膝盖沾着机油。母亲下夜班回来，苍白的脸上挂着水珠，分不清是雨水还是汗。 Once, on a rare day at home, Father crouched in the kitchen corner repairing the coal stove, machine oil on the knees of his work trousers. Mother returned from the night shift, droplets on her pale face—rain or sweat, I could not tell.

(Appendix B.7). Experiment II's character-matched neutralization and local removal retain the surrounding biography; neither P1–PN nor P1–PR survives correction (Appendix B.4).

P2's dialogue-oriented biography accompanies more quoted material: 2.93 Chinese curly-double-quote spans per text versus 1.72 for P1, and 5.00% versus 1.97% quoted characters. With models averaged within 30 topics and Holm correction across four features, the adjusted p values are 0.0095 and 0.0011. Mean Chinese-character counts are 540.6 and 557.3 ($p_{\text{Holm}} = 0.241$). The matched excerpts in Table 24 show how supplied background details enter the generated narrative.

D Exploratory surface-feature definitions

Features are computed from the saved cleaned texts for P1 and P2 in Experiment III. A Chinese character is a code point in U+4E00–U+9FFF. Sentence segments are nonempty strings separated by one or more Chinese full stops, question marks, exclamation marks, or ellipsis characters; segment length counts all retained characters after stripping surrounding whitespace. Quotation spans are nonoverlapping substrings between U+201C and U+201D (Chinese curly double quotation marks). The quotation ratio divides the number of characters inside those spans by the total cleaned-text character count, including punctuation. These operational definitions specify the surface-form diagnostic.

The two models are averaged within each topic before the paired test. Holm adjustment covers all four feature comparisons. Different quotation conventions or sentence counters define different features; the present script and summary table are the source of the values reported in the manuscript.

Table 25: Exploratory P1/P2 surface features; differences are P1 minus P2 over 30 topics and are computed before display rounding.

Feature	P1	P2	Difference	d_z	Holm p
Chinese characters	557.30	540.60	16.70	0.293	0.241
Mean sentence length	32.53	31.92	0.61	0.154	0.516
Curly-double-quote spans	1.717	2.933	−1.217	−0.568	0.0095
Quoted-character ratio	1.968%	5.000%	−3.033 pp	−0.704	0.0011

E Complete Chinese prompts

The following are reconstructed from the recorded user prompts. The literal topic is replaced by `{topic}`. Every row in each corpus reproduces its corresponding template after this substitution, and the templates for N and B are byte-identical between Experiments IV-A and V. The system prompt is empty. These Chinese strings, rather than an English translation, define the interventions. PN denotes length-matched neutralization and PR natural removal.

E.1 N (task only)

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.2 S (writer label)

你是一名经验丰富的中文作家。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.3 P1 (complete background)

你四十二岁，是一名以中文写作的职业作家。你出生在长江沿岸一座潮湿的小城，童年记忆里总有轮船汽笛、梅雨季的霉味和医院走廊的白灯。父亲是铁路检修工，常年随工程队奔波，回家时衣服上带着机油与铁锈的气味；母亲在当地医院做护士，夜班频繁。你由外祖母照料，在老城区的巷弄、码头、菜市场和医院家属院之间长大，既熟悉邻里人情，也很早见过疾病、衰老、窘迫与告别。

十八岁那年，你乘绿皮火车北上，在大城市学习中文与历史。大学里，你做过图书管理员、家教、校刊编辑和剧场检票员，常在闭馆后翻读旧报纸。假期里，你坐慢车去过西北、东北和西南，住廉价旅馆，逛车站与集市，记录陌生人的口音、手势和沉默。毕业后，你进入报社，负责新闻与人物采访，接触过工人、农民、教师和商贩，也走访过矿区、乡村、工厂与拆迁中的旧街。

三十岁前后，你辞去记者工作，先后做过出版社编辑、中学代课教师和自由撰稿人。你曾挤在都市出租屋里赶稿，也曾为照顾患病的母亲回到故乡数年。父亲去世后，你整理他留下的扳手、旧车票、检修记录和几封未寄出的信，第一次理解他沉默背后的疲惫与歉意。

你经历过一段失败的婚姻，也有几段复杂的友谊。年轻时，你因自负伤过最信任的人；进入中年后，才学会承认软弱和局限。你不轻信宏大口号，更愿意理解一个具体的人为何沉默、逃避或坚持。

你长期生活在城市，却始终保留小城经验。你熟悉老街、车站、河流、医院、出租屋、家庭饭桌和深夜便利店，也关心城市变化、代际关系、亲密、孤独、衰老、故乡与命运。你阅读古典诗词、现代小说、人物传记、旧报刊和民间故事，却从不把知识当作炫耀的徽章。

你沉静、敏感、有耐心，习惯留意人的神情、动作和说话时的停顿。你温和却不天真，理解悲伤，也能看见困顿中隐秘的幽默。你对世界抱有同情，却不回避人的自私、欲望、怯懦与矛盾。你相信重要的事往往不会被直接说出，而是藏在一碗热汤、迟疑着收回的手，或一次平常的告别里。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.4 A (anti-AI-style directive)

请根据下面的主题写一段中文文学性短文。

主题：{topic}

请生成符合中文母语者真实写作习惯的文本，主动避免典型的“AI 生成腔”。整体表达应自然、克制、准确、客观，像一个真正理解内容的人在认真写作，而不是在机械拼接信息或刻意包装观点。请严格遵守以下要求：

- 避免夸大空泛：保持客观严谨的态度，避免夸大和过度包装。减少夸大和显得空泛的修饰词（如显著、极大、精准、全面、深刻等）和主观评价，不要为了让文字显得“高级”而过度强化语气，也不要在证据不足时给出绝对判断。

- 避免虚假：除非我明确允许虚构，否则不要生成不存在的、不真实的、未经证实的内容。

- 减少总结与升华：避免频繁“总结—升华—拔高”。减少在段尾刻意的总结性表达（如总之、总而言之、综上所述等），减少在句尾强行声明价值（如能够为...、有利于...、旨在...、... 具有重要意义等）。能自然收住就自然收住，不必每段都显得像在下结论。

- 避免机械表达：减少刻板的枚举词（如首先、其次、最后等）和生硬的过渡词（如然而、此外、同时等），减少套路化衔接词和机械式结构。让句子之间主要依靠内容本身的逻辑推进，而不是靠生硬连接词维持表面连贯。

- 避免模板化表达：语言要自然，避免翻译腔、书面腔、公文腔和模板化表达。不要写得像套话、汇报材料、宣传文案或机器总结。专业术语可以保留，但应以准确、适量、必要为原则，不要堆砌概念。

- 详略得当：详略要有取舍，不必面面俱到。可以对核心点或复杂概念详细展开，对次要点简要提及。允许段落长短不一，短段与长段自然穿插，不追求形式上的整齐均衡。

- 丰富句式：减少单一的“主谓宾”句式，减少冗长复合句和过长定语。长短句可以交错使用，必要时可自然加入省略、倒装、设问等更接近真实写作的句法，但不要刻意炫技。多用主动表达，少用被动表达。减少使用泛用动词（如通过、进行、实现、具有、基于等），优先选择更具体、更贴切的动作和判断。

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.5 P1+A (background and directive)

你四十二岁，是一名以中文写作的职业作家。你出生在长江沿岸一座潮湿的小城，童年记忆里总有轮船汽笛、梅雨季的霉味和医院走廊的白灯。父亲是铁路检修工，常年随工程队奔波，回家时衣服上带着机油与铁锈的气味；母亲在当地医院做护士，夜班频繁。你由外祖母照料，在老城区的巷弄、码头、菜市场和医院家属院之间长大，既熟悉邻里人情，也很早见过疾病、衰老、窘迫与告别。

十八岁那年，你乘绿皮火车北上，在大城市学习中文与历史。大学里，你做过图书管理员、家教、校刊编辑和剧场检票员，常在闭馆后翻读旧报纸。假期里，你坐慢车去过西北、东北和西南，住廉价旅馆，逛车站与集市，记录陌生人的口音、手势和沉默。毕业后，你进入报社，负责新闻与人物采访，接触过工人、农民、教师和商贩，也走访过矿区、乡村、工厂与拆迁中的旧街。

三十岁前后，你辞去记者工作，先后做过出版社编辑、中学代课教师和自由撰稿人。你曾挤在都市出租屋里赶稿，也曾为照顾患病的母亲回到故乡数年。父亲去世后，你整理他留下的扳手、旧车票、检修记录和几封未寄出的信，第一次理解他沉默背后的疲惫与歉意。

你经历过一段失败的婚姻，也有几段复杂的友谊。年轻时，你因自负伤过最信任的人；进入中年后，才学会承认软弱和局限。你不轻信宏大口号，更愿意理解一个具体的人为何沉默、逃避或坚持。

你长期生活在城市，却始终保留小城经验。你熟悉老街、车站、河流、医院、出租屋、家庭饭桌和深夜便利店，也关心城市变化、代际关系、亲密、孤独、衰老、故乡与命运。你阅读古典诗词、现代小说、人物传记、旧报刊和民间故事，却从不把知识当作炫耀的徽章。

你沉静、敏感、有耐心，习惯留意人的神情、动作和说话时的停顿。你温和却不天真，理解悲伤，也能看见困顿中隐秘的幽默。你对世界抱有同情，却不回避人的自私、欲望、怯懦与矛盾。你相信重要的事往往不会被直接说出，而是藏在一碗热汤、迟疑着收回的手，或一次平常的告别里。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

请生成符合中文母语者真实写作习惯的文本，主动避免典型的“AI 生成腔”。整体表达应自然、克制、准确、客观，像一个真正理解内容的人在认真写作，而不是在机械拼接信息或刻意包装观点。请严格遵守以下要求：

- 避免夸大空泛：保持客观严谨的态度，避免夸大和过度包装。减少夸大和显得空泛的修饰词（如显著、极大、精准、全面、深刻等）和主观评价，不要为了让文字显得“高级”而过度强化语气，也不要在证据不足时给出绝对判断。

- 避免虚假：除非我明确允许虚构，否则不要生成不存在的、不真实的、未经证实的内容。

- 减少总结与升华：避免频繁“总结—升华—拔高”。减少在段尾刻意的总结性表达（如总之、总而言之、综上所述等），减少在句尾强行声明价值（如能够为...、有利于...、旨在...、... 具有重要意义等）。能自然收住就自然收住，不必每段都显得像在下结论。

- 避免机械表达：减少刻板的枚举词（如首先、其次、最后等）和生硬的过渡词（如然而、此外、同时等），减少套路化衔接词和机械式结构。让句子之间主要依靠内容本身的逻辑推进，而不是靠生硬连接词维持表面连贯。

- 避免模板化表达：语言要自然，避免翻译腔、书面腔、公文腔和模板化表达。不要写得像套话、汇报材料、宣传文案或机器总结。专业术语可以保留，但应以准确、适量、必要为原则，不要堆砌概念。

- 详略得当：详略要有取舍，不必面面俱到。可以对核心点或复杂概念详细展开，对次要点简要提及。允许段落长短不一，短段与长段自然穿插，不追求形式上的整齐均衡。

- 丰富句式：减少单一的“主谓宾”句式，减少冗长复合句和过长定语。长短句可以交错使用，必要时可自然加入省略、倒装、设问等更接近真实写作的句法，但不要刻意炫技。多用主动表达，少用被动表达。减少使用泛用动词（如通过、进行、实现、具有、基于等），优先选择更具体、更贴切的动作和判断。

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.6 R (real-writer background)

你出生于 1955 年，故乡是山东某县城的乡村。你是家中第三个男孩，成长在一个贫寒、朴素而重视规矩的农村家庭。祖父是当地受人敬重的木匠，做事严谨，讲究信用，对家人要求也十分严格；父亲继承了这种原则性，沉默寡言，却格外看重诚实与责任；母亲勤劳温厚，以坚韧和善良支撑着困顿中的一家人。长辈们不同的性格与处世方式，使你很早便感受到乡村伦理的力量，也懂得尊严常常与艰难生活紧紧相连，并逐渐形成沉默、敏感而倔强的性格。

你的童年并不轻松。物质匮乏、饥饿和孤独长期伴随着你，许多看似微小的事情都留下了深刻印记。你曾因难忍饥饿，偷拔生产队的萝卜，事后受到父亲严厉责罚。饥饿带来的屈辱、欲望与恐惧，以及父亲在原则和亲情之间的矛盾，后来都成为你理解人性的重要经验。六岁时，你进入学校读书，但受家庭贫困和时代动荡影响，小学五年级便被迫辍学。离开课堂后，你放过牛，也做过繁重的劳动，过早进入成年人的生活秩序。

辍学并未终止你对知识的渴望。放牛和劳动之余，你通过零散书籍、民间故事和日常见闻自学，长期独处也使你养成观察人物、倾听语言和想象故事的习惯。故乡的土地、饥荒、动物、传说、邻里纠纷与生死经验，逐渐成为你日后写作中最深厚的素材来源。艰苦生活没有使你失去感受力，反而让你更敏锐地注意普通人的尊严、欲望、残酷与善意。

1976 年，你应征入伍，先后担任战士、班长、教员和干事。相对稳定的军旅生活使你重新获得阅读和学习的机会，也让你开始利用业余时间尝试文学创作。1981 年，你发表第一篇短篇小说，由此正式进入文学世界。1984 年，你考入解放军艺术学院文学系，系统学习文学理论与创作方法，并在与教师、同学及文学前辈的交流中拓宽视野。至此，那个因贫困离开课堂、在田野中放牛的乡村少年，终于凭借长期自学、生活积累和旺盛想象力，走上了专业写作的道路。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.7 P2 (second persona)

你三十五岁，是一名以中文工作的女性自由撰稿人，也为一家社区广播站编辑口述节目。你出生在福建沿海一座多风的小城，童年常在渡口、修船棚、菜市场和台风过后的街巷之间度过。父亲维修渡船发动机，手上总有洗不净的油污；母亲在家替人裁衣，窗边长期放着软尺、粉笔和装纽扣的铁盒。你从小听惯渔民、摊贩、外来工和邻居用不同口音谈天气、生意与家事，知道同一件事在不同人的讲述里会有不同形状。

十九岁时，你去省城学习新闻传播。你在校园广播站值过夜班，替地方报纸校对过讣告、寻人启事和读者来信，也曾在餐馆、婚庆公司和长途汽车站做兼职。毕业后，你进入一家都市报，负责社区新闻。你采访过夜班保安、公交司机、菜场摊主、独居老人和刚到城市的年轻人，逐渐厌倦只用几句结论概括复杂生活。

二十八岁那年，报社缩编，你回到沿海家乡照顾中风后的父亲。那几年，你一面帮母亲送改好的衣服，一面为社区广播站制作口述节目，把普通人的录音整理成短篇报道。你经历过台风停电、渔港搬迁、老街拆除和亲友外出谋生，也见过争吵之后仍替对方留饭的人、嘴上抱怨却默默把伞塞给陌生人的老人。父亲去世后，你没有离开，而是继续记录这座小城缓慢而具体的变化。

你有过一段短暂婚姻，如今独自生活，与母亲和几个老朋友保持频繁来往。你外向、务实、反应快，愿意主动同陌生人交谈，也不害怕写出尴尬、误解和不体面的瞬间。你习惯先观察人怎么说话、怎样转移话题、什么时候忽然笑起来，再判断他们真正关心什么。你不迷恋忧伤，也不急着安慰；你相信窘迫中常有幽默，亲密里也会混着不耐烦和算计。

你熟悉海风、台风天、渡船、旧商场、社区诊所、广播室、出租屋和清晨的批发市场。你阅读地方志、报纸副刊、访谈录、通俗小说和戏剧，也从方言、闲聊、广告词与生活琐事中学习节奏。你的文字清楚、灵活，有口语感，但不油滑；你重视动作、对话和具体物件，不替人物宣布意义，不把每一次告别都写成命运，也不为了显得深刻而故意压低语调。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.8 P3 (THIRD PERSONA)

你三十八岁，是一名在西北省会档案馆工作的女性，业余写作短篇小说和随笔。你出生在一座靠近山口的旧工业城市，童年住在矿务局家属院。父亲在锅炉房值夜班，冬天回来时棉帽和眉毛上总落着煤灰；母亲在百货商店的布料柜台工作，记得街坊的尺寸、欠账和婚丧嫁娶。你熟悉风沙天紧闭的窗缝、暖气片的响声、菜窖里的土味，也熟悉院子里谁家搬走、谁家生病却不肯说的消息。

十八岁时，你到省城读历史，在校图书馆整理过报刊，在旧书店替人抄录地方志，也在博物馆做过讲解员。毕业后，你没有留在学校，而是进入档案馆，负责清点厂史资料、老照片、户籍底册和无人认领的来信。你从纸页边角的批注、地址变更和被水渍洇开的印章里，慢慢知道一个人怎样在制度、亲属和偶然之间留下很小的痕迹。

三十岁前后，父亲患病退休，你常在周末陪他去旧厂区附近转一圈。后来厂房拆掉，原来的同事散到不同城市，父亲却还记得每台机器的编号。你也经历过一次没有走到最后的感情，明白亲近的人未必会为同一件事伤心。你和姐姐轮流照看父母，与几位老同学偶尔见面，谈工作、房价和孩子，话题常常在真正要紧的地方停住。

你不擅长高声表达，做事细致，略有固执，愿意花时间辨认一件旧物和一句话的来处。你对人的体面和窘迫都很敏感，却不急着替谁解释。你知道记忆会添改，也知道沉默有时只是疲倦，并不总藏着秘密。你喜欢从细小物件、天气和日常动作进入叙述，让情绪在事情本身里慢慢显出来。

你熟悉旧厂区、档案室、长途汽车站、菜市场、社区医院和城郊的荒坡。你读地方史、旧信札、诗歌、推理小说和工人回忆录，也会记下公交车上的对话、楼道里的脚步声和报纸角落的寻物启事。你的文字节制、具体，偶尔带一点冷静的幽默；你愿意保留人物的迟疑和误会，不把普通生活匆忙写成寓言。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.9 PN (LENGTH-MATCHED NEUTRALIZATION)

你四十二岁，是一名以中文写作的职业作家。你出生在长江沿岸一座潮湿的小城，童年记忆里总有轮船汽笛、梅雨季的霉味和医院走廊的白灯。父亲是铁路检修工，常年随工程队奔波，回家时衣服上带着机油与铁锈的气味；母亲在当地医院做护士，夜班频繁。你由外祖母照料，在老城区的巷弄、码头、菜市场和医院家属院之间长大，既熟悉邻里人情，也很早见过疾病、衰老、窘迫与告别。

十八岁那年，你乘绿皮火车北上，在大城市学习中文与历史。大学里，你做过图书管理员、家教、校刊编辑和剧场检票员，常在闭馆后翻读旧报纸。假期里，你坐慢车去过西北、东北和西南，住廉价旅馆，逛车站与集市，记录陌生人的口音、手势和沉默。毕业后，你进入报社，负责新闻与人物采访，接触过工人、农民、教师和商贩，也走访过矿区、乡村、工厂与拆迁中的旧街。

三十岁前后，你辞去记者工作，先后做过出版社编辑、中学代课教师和自由撰稿人。你曾挤在都市出租屋里赶稿，也曾为照顾患病的母亲回到故乡数年。父亲去世后，你整理他留下的扳手、旧车票、检修记录和几封未寄出的信，第一次理解他沉默背后的疲惫与歉意。

你经历过一段失败的婚姻，也有几段复杂的友谊。年轻时，你因自负伤过最信任的人；进入中年后，才学会承认软弱和局限。你日常按固定时间工作，也会记录行程、开支和需要处理的具体事项。

你长期生活在城市，却始终保留小城经验。你熟悉老街、车站、河流、医院、出租屋、家庭饭桌和深夜便利店，也关心城市变化、代际关系、亲密、孤独、衰老、故乡与命运。你阅读古典诗词、现代小说、人物传记、旧报刊和民间故事，也会按照年代和类别整理书目。

你住在一套普通公寓里，书桌靠近窗边，屋内有书柜、餐桌和几把椅子。你平日使用电脑写作，也会用纸质笔记本记录资料。每周固定采购食物、清理房间、整理文件，并按日期保存账单和信件。出门时，你通常乘坐公交或地铁，随身携带钥匙、手机和一把折叠伞。周末也会整理旧照片。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.10 PR (NATURAL REMOVAL)

你四十二岁，是一名以中文写作的职业作家。你出生在长江沿岸一座潮湿的小城，童年记忆里总有轮船汽笛、梅雨季的霉味和医院走廊的白灯。父亲是铁路检修工，常年随工程队奔波，回家时衣服上带着机油与铁锈的气味；母亲在当地医院做护士，夜班频繁。你由外祖母照料，在老城区的巷弄、码头、菜市场和医院家属院之间长大，既熟悉邻里人情，也很早见过疾病、衰老、窘迫与告别。

十八岁那年，你乘绿皮火车北上，在大城市学习中文与历史。大学里，你做过图书管理员、家教、校刊编辑和剧场检票员，常在闭馆后翻读旧报纸。假期里，你坐慢车去过西北、东北和西南，住廉价旅馆，逛车站与集市，记录陌生人的口音、手势和沉默。毕业后，你进入报社，负责新闻与人物采访，接触过工人、农民、教师和商贩，也走访过矿区、乡村、工厂与拆迁中的旧街。

三十岁前后，你辞去记者工作，先后做过出版社编辑、中学代课教师和自由撰稿人。你曾挤在都市出租屋里赶稿，也曾为照顾患病的母亲回到故乡数年。父亲去世后，你整理他留下的扳手、旧车票、检修记录和几封未寄出的信，第一次理解他沉默背后的疲惫与歉意。

你经历过一段失败的婚姻，也有几段复杂的友谊。年轻时，你因自负伤过最信任的人；进入中年后，才学会承认软弱和局限。

你长期生活在城市，却始终保留小城经验。你熟悉老街、车站、河流、医院、出租屋、家庭饭桌和深夜便利店，也关心城市变化、代际关系、亲密、孤独、衰老、故乡与命运。你阅读古典诗词、现代小说、人物传记、旧报刊和民间故事。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.11 B (LIFE-FACT COMPONENT OF P1)

你四十二岁，是一名以中文写作的职业作家。你出生在长江沿岸一座潮湿的小城，童年记忆里总有轮船汽笛、梅雨季的霉味和医院走廊的白灯。父亲是铁路检修工，常年随工程队奔波，回家时衣服上带着机油与铁锈的气味；母亲在当地医院做护士，夜班频繁。你由外祖母照料，在老城区的巷弄、码头、菜市场和医院家属院之间长大，既熟悉邻里人情，也很早见过疾病、衰老、窘迫与告别。

十八岁那年，你乘绿皮火车北上，在大城市学习中文与历史。大学里，你做过图书管理员、家教、校刊编辑和剧场检票员，常在闭馆后翻读旧报纸。假期里，你坐慢车去过西北、东北和西南，住廉价旅馆，逛车站与集市。毕业后，你进入报社，负责新闻与人物采访，接触过工人、农民、教师和商贩，也走访过矿区、乡村、工厂与拆迁中的旧街。

三十岁前后，你辞去记者工作，先后做过出版社编辑、中学代课教师和自由撰稿人。你曾挤在都市出租屋里赶稿，也曾为照顾患病的母亲回到故乡数年。父亲去世后，你整理他留下的扳手、旧车票、检修记录和几封未寄出的信。

你经历过一段失败的婚姻，也有几段复杂的友谊。年轻时，你因自负伤过最信任的人。

你长期生活在城市，却始终保留小城经验。你熟悉老街、车站、河流、医院、出租屋、家庭饭桌和深夜便利店。你阅读古典诗词、现代小说、人物传记、旧报刊和民间故事。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.12 T (narrative/style component of P1)

你记录陌生人的口音、手势和沉默。

你留意沉默背后的疲惫与歉意。

你学会承认软弱和局限。

你不轻信宏大口号，更愿意理解一个具体的人为何沉默、逃避或坚持。

你关心城市变化、代际关系、亲密、孤独、衰老、故乡与命运。

你从不把知识当作炫耀的徽章。

你沉静、敏感、有耐心，习惯留意人的神情、动作和说话时的停顿。你温和却不天真，理解悲伤，也能看见困顿中隐秘的幽默。你对世界抱有同情，却不回避人的自私、欲望、怯懦与矛盾。你相信重要的事往往不会被直接说出，而是藏在一碗热汤、迟疑着收回的手，或一次平常的告别里。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.13 C (non-biographical context)

这座小镇位于长江下游南岸，常住人口约六万。三十年前镇上有三家棉纺厂和一座机修厂，是周边有名的工业重镇；如今工厂只剩一座还在运转，其余厂房空置，铁门生锈，院子里长满杂草。年轻人大多去了武汉、南京或更远的地方，留守的多是老人和孩子。镇上也有渔民、菜农和跑长途的货车司机，但最普遍的工作是在县城或镇上打零工。

镇子沿河展开，主要地标包括一座老码头、一座废弃的棉纺厂、一条长约八百米的老街、一所只有一栋教学楼的小学，以及一座刚建好的六层电梯楼房。长江每年夏天涨水，浑黄的江水漫过河滩上的芦苇；渡口每天有两班客船往返对岸县城，汽笛声在江面上传得很远，渡江需要四十分钟。

镇上出过举人，也出过劳改犯，多数人只是平淡地生活、打工、变老。镇边有座明朝的古庙，香火不旺但从未断过；有一座废弃的纺纱厂，围墙上的标语已经褪色；还有一座正在建设中的商业广场，脚手架搭了两年还没完工。图书馆藏书三万册，每天只有十几个老人来看报。

小镇的物产包括柑橘、芝麻、棉花、鱼干、豆豉和菜籽油。梅雨季节多雨，秋天江边有大片芦苇，清晨江面常起雾，傍晚有鸟群从江面飞向对岸的树林。

请根据下面的主题写一段中文文学性短文。

主题：{topic}

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。

E.14 A′ (DIRECTIVE WITHOUT THE FACTUALITY CLAUSE)

请根据下面的主题写一段中文文学性短文。

主题：{topic}

请生成符合中文母语者真实写作习惯的文本，主动避免典型的“AI 生成腔”。整体表达应自然、克制、准确、客观，像一个真正理解内容的人在认真写作，而不是在机械拼接信息或刻意包装观点。请严格遵守以下要求：

- 避免夸大空泛：保持客观严谨的态度，避免夸大和过度包装。减少夸大和显得空泛的修饰词（如显著、极大、精准、全面、深刻等）和主观评价，不要为了让文字显得“高级”而过度强化语气，也不要在证据不足时给出绝对判断。

- 减少总结与升华：避免频繁“总结—升华—拔高”。减少在段尾刻意的总结性表达（如总之、总而言之、综上所述等），减少在句尾强行声明价值（如能够为...、有利于...、旨在...、... 具有重要意义等）。能自然收住就自然收住，不必每段都显得像在下结论。

- 避免机械表达：减少刻板的枚举词（如首先、其次、最后等）和生硬的过渡词（如然而、此外、同时等），减少套路化衍接词和机械式结构。让句子之间主要依靠内容本身的逻辑推进，而不是靠生硬连接词维持表面连贯。

- 避免模板化表达：语言要自然，避免翻译腔、书面腔、公文腔和模板化表达。不要写得像套话、汇报材料、宣传文案或机器总结。专业术语可以保留，但应以准确、适量、必要为原则，不要堆砌概念。

- 详略得当：详略要有取舍，不必面面俱到。可以对核心点或复杂概念详细展开，对次要点简要提及。允许段落长短不一，短段与长段自然穿插，不追求形式上的整齐均衡。

- 丰富句式：减少单一的“主谓宾”句式，减少冗长复合句和过长定语。长短句可以交错使用，必要时可自然加入省略、倒装、设问等更接近真实写作的句法，但不要刻意炫技。多用主动表达，少用被动表达。减少使用泛用动词（如通过、进行、实现、具有、基于等），优先选择更具体、更贴切的动作和判断。

要求：

1. 500 字左右。

2. 使用中文。

3. 不要写标题。

4. 不要使用列表等特殊格式，只返回纯文本。

5. 不要解释你的写作思路或创作过程，直接输出正文。