Grammar Correction and Phrasing Refinement Lower Sentence-Level Perplexity and Entropy of Human Formal Writing

Nabil Islam[1*]

[1*]Independent Researcher.

Corresponding author(s). E-mail(s): w3nabil@gmail.com;

Abstract

Many AI-text detectors rely on perplexity, that is, on how predictable a text is to a language model. A natural question is whether a human can produce perplexity and entropy values similar to those of AI-generated text. We conduct a preliminary sentence-level comparison of five writing conditions: original human-written text, grammar-corrected human text, phrasing-refined human text, text from GLM 5.3 prompted to mimic the human author, and text generated by GLM 5.3 Flash without any human sample. All 13 matched sentences were scored with Qwen3. The mean perplexity decreased from 39.59 for the original human response to 37.24 after grammar correction and 21.26 after phrasing refinement, compared with 26.92 for the GLM 5.3 mimicking response and 37.52 for the GLM 5.3 Flash response. Mean entropy (in bits) was highest for the original and grammar-corrected human texts (2.98 and 3.05) and for the AI-generated texts it was 2.59 and 2.73, while the human-refined text had the lowest value (2.25). Sentence-to-sentence entropy changes were negatively autocorrelated for all human-originated texts and positively autocorrelated for both AI-generated texts, although none of the AR(1) coefficients was statistically significant. These results indicate that a polished human-written formal response can reach lower perplexity and entropy than AI-generated text, which may lead to false positives in perplexity-based detection.

Keywords: AI-generated text detection, perplexity, Shannon entropy, human writing, false positives

1 Introduction

Large language models (LLMs) are now widely used for writing, editing and text generation, and this has increased the interest in finding statistical characteristics that separate human-authored text from machine-generated text. One of the most common suggestions is perplexity, which measures how predictable a sequence of tokens is under a language model [1, 2]. The intuition is simple: text sampled from a language model tends to be probable under a language model, so it should receive a lower perplexity than text written by a person. A recent large-scale benchmark reports that AI-generated responses indeed tend to have lower perplexity than human responses [3]. This is a useful direction because it could allow AI-generated text to be tested statistically, instead of depending on paid services from commercial vendors.

There are, however, reasons for caution. The effectiveness of perplexity-based detection depends on the scoring model, the data distribution and the detection procedure [2, 4]. Detectors have been shown to misclassify the writing of non-native English speakers as AI-generated, mainly because their text has lower perplexity [4], and independent testing of commercial detectors found them neither accurate nor reliable enough for high-stakes decisions [5]. A natural question follows: if a human is able to produce lower perplexity like AI-generated text, what makes a human text more or less predictable? This is the question we try to look at here, using a deliberately simple and transparent set-up.

Shannon entropy [6] gives a second direction. Taking the entropy of the model's predictive distribution measures how uncertain the model is about what comes next, which is related to, but not the same thing as, perplexity. A recent study states that human text generally has higher average entropy, while machine-generated text has sharper local entropy fluctuations [7]. We therefore calculate sentence-level entropy for each of our samples to check whether any statistical pattern separates the AI-generated text from the human text.

1.1 Related work

Statistical and zero-shot detection.

Early statistical detectors such as GLTR used the rank and probability of each token under a language model to highlight text that looks machine-generated [8]. Later zero-shot methods compare the log-probability of a text with that of perturbed or conditionally resampled versions of it [9, 10], or contrast the scores of two related models [11]. All of these build on the same basic signal, namely how likely the text is under a language model. Ippolito et al. [12] showed that the decoding strategy changes detectability, with the strategies that produce the most fluent text being the hardest for humans to spot but the easiest for automatic detectors. This suggests that lower perplexity is often a feature of fluent, high-probability text in general.

Human text is not maximally probable.

Holtzman et al. [13] observed that human text does not sit at the high-probability end of the model distribution and that the per-token probability of human text fluctuates in a way that likelihood-maximising decoding does not reproduce. This is the reason

why perplexity and entropy are expected to differ between human and machine text. It also means that the difference should shrink when a human text is edited towards fluent, expected wording.

Limits of detection.

Paraphrasing is known to reduce the accuracy of several detectors considerably [14], and there are theoretical arguments that, as language models get closer to human text, even the best possible detector becomes only slightly better than a random classifier [15]. These works mostly study machine text that has been modified to evade detection. The opposite situation, in which human text is edited and then possibly judged to be machine-like, has received much less attention.

Entropy over a discourse.

At a longer range, it has been argued that the entropy of text tends to be roughly constant per sentence once the context is taken into account [16], which makes the sentence-to-sentence behaviour of entropy a reasonable thing to look at. Wang et al. [7] make use of this idea by working with local entropy fluctuations for black-box detection.

1.2 Contributions

Previous work has examined machine-generated text detection and the effects of paraphrasing on detection performance, but it rarely discusses why human writing is sometimes as predictable as AI writing, or how much the editing of an otherwise human text changes its statistics. Grammatical correction or stylistic revision does not change who the author is, but it can still change the lexical choices, the syntactic structures and the token probabilities. A change in perplexity after editing therefore does not mean that the text became more or less human. Our contributions are:

1. Human formal writing can reach AI-like perplexity and can even go below it when awkward phrasings are replaced with natural expressions.
2. Grammar correction lowers the mean perplexity only slightly, while phrasing refinement lowers it dramatically.
3. The original and grammar-corrected human texts have a higher mean entropy and the highest entropy value than both AI-generated texts, whereas the human-refined text has the lowest entropy of all five conditions.
4. All three human-originated texts have a negative first-order autoregressive (AR(1)) coefficient on the sentence entropy sequence, while both AI-generated texts have a positive one. The coefficients are not statistically significant in our small sample.

2 Data and Methods

2.1 Writing conditions

We separate our samples into two categories, human-originated and AI-generated. The human-originated category has three versions of the same response:

- **Human-written**: the response where a human independently answered a question.
- **Grammar-corrected**: the human-written response with possible grammar errors fixed and nothing else changed.
- **Human-refined**: the human-written response where an expert replaced awkward phrasings and unnatural flow with more natural expressions in order to polish the writing.

All three are written, fixed or modified by humans. The AI-generated category answers the same question using GLM 5.3 and GLM 5.3 Flash, in two variants:

- **GLM 5.3 Mimicking**: the grammar-corrected sample was given to GLM 5.3 as a writing sample, and the model was prompted to adopt and mimic the human author while keeping the original tone, habits and so on.
- **GLM-5.3-Flash**: raw writing from GLM 5.3 Flash with no human writing sample.

All five responses answer the same IELTS Task 2 question, "Some people believe that university education should be free for everyone, regardless of their financial background. To what extent do you agree or disagree?", with an IELTS examiner as the intended audience. The grammar-corrected human essay has 300 words. Both AI models were asked for exactly 13 sentences and 250–350 words, and they produced 275 words (GLM 5.3 Mimicking) and 334 words (GLM-5.3-Flash). Each response consists of 13 sentences, and the sentences were matched across the five conditions by position. The GLM 5.3 Mimicking prompt contained the grammar-corrected human essay as the writing sample, whereas the GLM-5.3-Flash prompt contained only the task and the constraints, with an instruction not to copy from memory or from any online source. Both prompts are given in Appendix D.

2.2 Scoring

Each response was scored with Qwen3 [17], run as a quantised GGUF model through llama.cpp. The model was the 8B instruction-tuned version of Qwen3 (a community "AWQ-compatible" re-release), quantised to 4 bits with the Q4_K_M scheme. The texts were scored as plain text, without a chat template. Text was tokenised with the model's own tokenizer, and a start-of-text token was placed before the first token. Each response was scored in a single pass (a sliding window of up to 1024 tokens is used for longer texts), so every token is conditioned on all of the preceding text of the same response, and sentence-level values come from grouping the tokens of each sentence, written one sentence per line. *Perplexity* (PPL) of a sentence of N tokens is the exponential of the mean negative log-likelihood per token,

$$\mathrm{PPL} = \exp\left(-\frac{1}{N}\sum_{j=1}^{N}\log p(x_j \mid x_{<j})\right), \tag{1}$$

where a lower value means the model finds the sentence more predictable. *Entropy* is the Shannon entropy in bits [6], $H_j = -\sum_v p(v \mid x_{<j}) \log_2 p(v \mid x_{<j})$, of the model's next-token distribution, averaged over the tokens in the sentence. It is computed from the full softmax over the vocabulary, so it describes the whole predictive distribution

and not only the probability of the token that was actually written. In the code, PPL is obtained as $2^{\bar{s}}$ with $\bar{s}$ the mean surprisal in bits, which is the same quantity as the equation above. Because the first sentence of each response is scored with only the start token as context, it behaves differently from the others (it has the highest PPL in four of the five conditions), so we also report an adjusted mean that excludes it.

2.3 Local fluctuation measures

To look at how entropy changes along the response, we use three simple time-series measures on the 13-sentence entropy sequence $y_1, \ldots, y_{13}$. Let y_t be the entropy of sentence t.

Sentence-to-sentence net change.

The first-order difference between consecutive sentences is

$$\Delta y_t = y_t - y_{t-1}, \qquad t \in \{2, 3, \ldots, 13\}. \tag{2}$$

We also count how often the sign of Δy_t flips between consecutive transitions (11 possible flips).

Autoregressive model, AR(1).

To test whether the sentence transitions follow a predictable linear process, we fit

$$y_t = c + \phi\, y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{3}$$

by ordinary least squares [18], where c is a constant, ϕ is the lag-1 coefficient measuring directional momentum, and ε_t is white noise. A negative ϕ means that a high-entropy sentence tends to be followed by a lower one (mean-reverting, zig-zag behaviour), and a positive ϕ means that the entropy tends to continue in the same direction.

Rolling local variance.

To capture the stability of the local variance, we compute a 3-sentence rolling sample variance,

$$S_t^2 = \frac{1}{k-1} \sum_{i=0}^{k-1} (y_{t-i} - \bar{y}_t)^2, \qquad k = 3, \tag{4}$$

where $\bar{y}_t$ is the mean of the window, and summarise it with the Variance Spread Ratio,

$$\text{Variance Spread} = \frac{\max_t \left(S_t^2\right)}{\min_t \left(S_t^2\right)}. \tag{5}$$

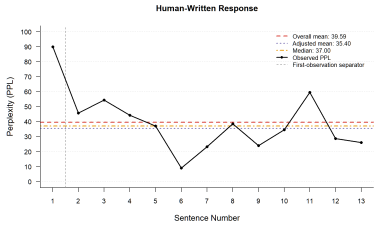
2.4 Paired comparison

As a rough check on the size of the differences, we compared the original human-written response with each of the other four conditions using a two-sided Wilcoxon signed-rank test on the 13 matched sentences. These tests are exploratory and uncorrected

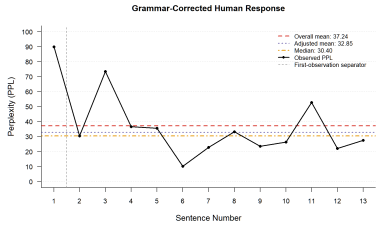
for multiple comparisons, and since all sentences come from a single essay they are not independent samples, so the p-values should be read as descriptive.

3 Results

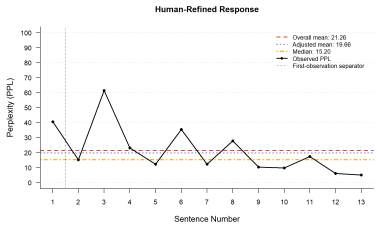
3.1 Sentence-level perplexity



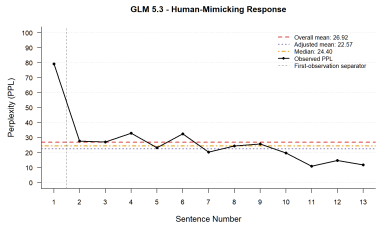
(a) Human-written response



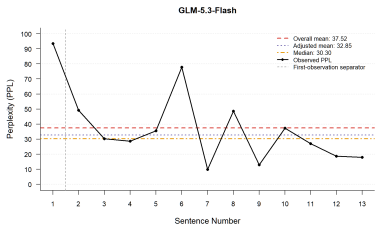
(b) Grammar-corrected human response



(c) Human-refined response



(d) GLM 5.3 human-mimicking response



(e) GLM-5.3-Flash

Fig. 1: Sentence-level perplexity (PPL) trajectories across five writing conditions evaluated using Qwen3. The solid black lines represent observed sentence-level PPL scores, while the horizontal reference lines represent the overall mean, adjusted mean, and median for each condition. The adjusted mean excludes the first sentence.

As visualised in Figure 1 and summarised in Table 1, the raw human-written response had a mean sentence-level perplexity of approximately 39.59 under Qwen3.

After grammatical correction the mean decreased to 37.24, a reduction of about 2.35 points. This is a small change, which suggests that grammatical correction alone has a limited effect on the overall PPL distribution (paired Wilcoxon $p = 0.13$). The sentence-level results also show that corrections do not reduce PPL uniformly: Sentence 2 decreased from 45.7 to 30.4, whereas Sentence 3 increased from 54.4 to 73.5, and Sentences 6 and 13 had slight increases (9.0 to 10.1 and 26.0 to 27.5). Eight sentences became lower, three became higher and the first sentence was unchanged. Grammatical correction can therefore push the predictability of a sentence in either direction, depending on the original construction and the modification, although the overall tendency is a modest decrease.

Table 1: Summary of sentence-level PPL and entropy (bits) across the 13 matched sentences. The adjusted mean excludes the first sentence; SD is the sample standard deviation.

	PPL				**Entropy (bits)**			
Condition	Mean	Adj.	Median	SD	Mean	Adj.	Median	SD
Human-Written	39.59	35.40	37.00	20.39	2.975	2.926	3.010	0.351
Grammar-Corrected	37.24	32.85	30.40	22.31	3.045	3.002	3.120	0.373
Human-Refined	21.26	19.66	15.20	16.31	2.249	2.195	2.180	0.334
GLM 5.3 Mimicking	26.92	22.58	24.40	17.18	2.592	2.527	2.540	0.454
GLM-5.3-Flash	37.52	32.85	30.30	24.81	2.731	2.732	2.720	0.402

A much larger reduction was observed when the human-written response was refined, with unnatural or awkward expressions replaced by more natural alternatives. The refined response reached a mean PPL of about 21.26, more than 18 points below the original (Wilcoxon $p = 0.013$). Eleven of the 13 sentences had lower PPL after refinement; the exceptions were Sentences 3 and 6, which increased from 54.4 to 61.4 and from 9.0 to 35.4. The effect is therefore not universal, but its size suggests that phrasing and linguistic naturalness can influence PPL far more than grammar alone.

The AI-generated conditions add more context. The GLM 5.3 mimicking response had a mean PPL of about 26.92, which is 5.66 points *higher* than the human-refined response. The refined human text was thus, on average, more predictable to Qwen3 than an AI text that was explicitly written to imitate the same author. This is at odds with the assumption that a low PPL means AI authorship, and it supports the idea that naturally phrased human writing can have a lower PPL than AI text, even when the AI text is built to mimic a human. The mimicking response was also significantly lower than the original human text (Wilcoxon $p = 0.017$), which fits the expectation that a model asked to imitate a person produces smoother text than the person did.

GLM-5.3-Flash showed a different pattern. Its mean PPL of 37.52 is far above that of the mimicking response and nearly equal to the grammar-corrected response (37.24), and it is also close to the raw human text (39.59; Wilcoxon $p = 0.27$). Similar

means should not be read as similar sentence-level behaviour, however. GLM-5.3-Flash had the largest standard deviation (24.81) and particularly high values at Sentences 1 and 6 (93.6 and 77.9). The mean alone does not describe these distributions well.

Taken together, sentence-level PPL is sensitive to how the language is formulated, and is not determined by authorship alone. Grammar correction gave a small reduction, phrasing refinement a large one, and the two AI-generated conditions differed considerably from each other, so AI text is not uniformly low in PPL. The human-refined response had a lower mean PPL than both AI-generated responses, while GLM-5.3-Flash stayed close to the human-originated samples. A low PPL cannot therefore be taken as proof of AI-generated content. Perplexity-based judgements should take into account linguistic refinement, sentence-level variation and the language model used for scoring. All sentence-level values are given in Appendix Table A1.

3.2 Sentence-level entropy

The entropy values from Qwen3 also differ across the five conditions (Table 1). The human-refined response has the lowest mean entropy, 2.249 bits, against 2.975 bits for the original human-written response, which is a reduction of about 24.4% in predictive uncertainty (Wilcoxon $p < 0.001$). In contrast, the grammar-corrected response has a slightly higher mean entropy of 3.045 bits (median 3.120), so grammar correction does not necessarily reduce the model's uncertainty. Both AI-generated responses lie in between, at 2.592 bits for GLM 5.3 Mimicking and 2.731 bits for GLM-5.3-Flash. The highest single value, 3.56 bits, occurs at Sentence 1 of the human-written and grammar-corrected responses.

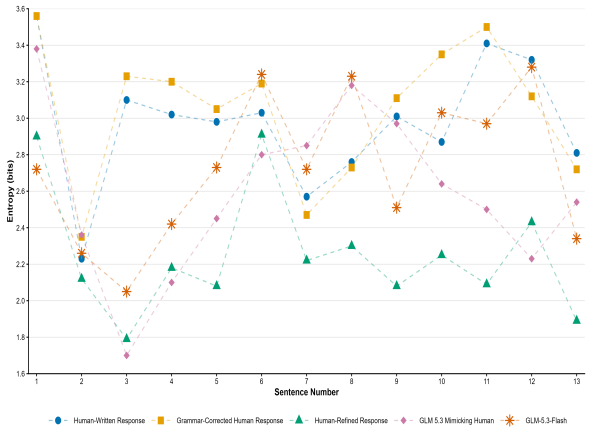


Fig. 2: Sentence-level entropy trajectories across the five writing conditions, measured using Qwen3. Distinct marker shapes represent individual writing conditions, while dashed lines illustrate changes in entropy between consecutive sentences.

At the sentence level, the human-refined response has a lower entropy than the original in all 13 sentences. The two AI-generated responses show more varied trajectories, and fluctuate differently across the sequence (Figure 2). Linguistic refinement therefore appears to affect predictive uncertainty more consistently than grammar correction, and AI-generated text does not show uniformly low entropy.

Overall, entropy is sensitive to linguistic transformations, in particular to phrasing refinement. That human-authored text can become considerably more predictable to a language model without any change of authorship means that entropy should not be read as a direct indicator of AI authorship. It is better seen as a measure of predictive uncertainty that adds some information about the statistical character of a text.

3.3 Local entropy fluctuations

Table C3 (Appendix) lists the sentence-to-sentence net changes. Counting sign flips between consecutive transitions gives 8 of 11 for the human-written text, 6 of 11 for the grammar-corrected text, 10 of 11 for the human-refined text, 3 of 11 for GLM 5.3 Mimicking and 8 of 11 for GLM-5.3-Flash. The human-refined text therefore alternates up and down most often, and the mimicking response shows the longest runs in one direction, but GLM-5.3-Flash alternates about as often as the original human text, so this measure alone does not separate the two sources.

Table 2: AR(1) and local rolling variance metrics of the sentence entropy sequence across the five conditions. p-values are two-sided, from the OLS slope t-test with 11 degrees of freedom.

Category	**AR(1)** ϕ	p**-value**	**Max Var**	**Min Var**	**Var Spread Ratio**
Human-Written	-0.286	0.299	0.456	0.0007	651.76×
Grammar-Corrected	-0.126	0.678	0.391	0.0070	55.63×
Human-Refined	-0.164	0.551	0.325	0.0091	35.70×
GLM 5.3 Mimicking	$+0.405$	0.120	0.716	0.0279	25.68×
GLM-5.3-Flash	$+0.190$	0.571	0.229	0.0270	8.49×

The AR(1) results are given in Table 2. All three human-originated texts have a negative ϕ, which indicates a mean-reverting, zig-zag pattern: if sentence N has a high entropy, sentence $N+1$ is likely to have a lower one. This is plausible for human writing, which is driven by sudden changes in thought, structure and tone. A dense, complex sentence (high entropy) is often followed by a short, simple one (low entropy), and this back-and-forth gives a negative lag coefficient. Both AI-generated texts have a positive ϕ. A language model generates text token by token from a continuous context, so the distribution of one sentence carries over smoothly into the next, which could create the momentum, or positive autocorrelation, that we see. We stress, however, that none of the coefficients is statistically significant ($p \geq 0.12$), and with 12 transition pairs per condition the sign of ϕ is a tentative observation and not a finding.

The variance spread ratio of the human-originated texts (651.76, 55.63 and 35.70) is larger than that of the AI-generated texts (25.68 and 8.49), meaning that the local

variance of human entropy is less uniform along the response. The ratio is, however, very sensitive to the minimum, and the original human text has a nearly flat stretch (minimum variance 0.0007) that dominates the number. The maximum local variance is actually the highest for GLM 5.3 Mimicking (0.716). This does not straightforwardly match the claim that machine text shows sharper local entropy fluctuations [7], and the difference may come from the short sequences, a single topic and the use of a human-imitating prompt.

4 Discussion

The main observation is that how a human text is edited can move its perplexity and entropy across the range occupied by AI text. Grammar correction alone changed the statistics very little. Replacing awkward phrasing with natural wording moved the mean PPL from 39.59 to 21.26, below both AI conditions, and moved the mean entropy from 2.98 to 2.25 bits. This agrees with what is known about the relation between fluency and probability: fluent, conventional wording is high-probability wording [12, 13], and removing the unusual phrasing of a non-expert writer makes the text more predictable. It also fits the finding that perplexity-based detectors disadvantage writers with a limited expression range [4]. A person who writes in a polished, conventional formal register may then be at risk of being flagged, and the more careful the editing, the higher that risk may be.

The entropy results point the opposite way for grammar correction: entropy slightly increased, even though PPL slightly decreased. PPL and entropy are related but measure different things, since PPL evaluates the probability of the words that were actually written and entropy describes the spread of the model's whole predictive distribution. A text can therefore be easy to predict at the chosen tokens while the model remains uncertain about alternatives. Using both measures together is more informative than either alone.

Finally, the sign of the AR(1) coefficient, and to some extent the sign-flip counts, give a hint of a "breathing" pattern in human-originated texts and a more momentum-like pattern in the human-mimicking AI text. GLM-5.3-Flash does not follow this clearly, and we cannot claim that fluctuation measures separate the two sources. A larger study is needed before this is more than a hypothesis.

5 Limitations

This study is preliminary and has several limitations.

First, the observations come from one human-written response, so the results may reflect the specific essay, author or topic. A larger dataset with multiple essays and authors is required to assess generalisability, and expanding the study in this way is our future work. The sentences within an essay are also not independent, which weakens the Wilcoxon tests.

Second, only one scoring model (Qwen3 8B, 4-bit quantised) was used, and it is an instruction-tuned model. Such models are tuned towards fluent assistant-style text, so their perplexity and entropy values may differ from those of a base model, and repeating the analysis with a base model is a natural next step. Numerical results

and paired statistical analyses across several evaluator models are needed to establish whether the size of the effects is consistent across evaluators. Only two AI models, from one family, were compared, and AI-generated text may behave differently with other models, prompts and decoding settings.

Third, the study does not include independent human judgments or validated AI-detector outputs. It therefore measures changes in model-assigned perplexity and entropy, not changes in actual human-likeness or in the probability of being flagged by a detector.

Fourth, with 13 sentences per condition, the AR(1) estimates and the variance spread ratios are unstable, and none of the AR(1) coefficients is significant. The ratio in particular depends on a single minimum value.

Finally, each sentence is scored in the context of the sentences before it, so the sentence-level values are not independent of each other, and an edit in one sentence can change the scores of the sentences that follow. This is closer to how a detector sees a whole document, but it also means that the autocorrelation measures reflect both the writing and the model's conditioning on earlier text. The scores are tied to one tokenizer, one quantised model and one context setting, and sentence boundaries were defined by line breaks.

6 Conclusion

In this preliminary study, a polished human-written formal response reached a lower perplexity and, after refinement, a lower entropy than the AI-generated texts we compared it with. Grammar correction had a small effect on perplexity and almost none on entropy, whereas phrasing refinement had a large effect on both. Perplexity alone therefore cannot be used to decide whether a text is AI-generated, and a low value may simply reflect careful editing. It may still give an optional level of direction when it is combined with other evidence. The sentence-level trajectories of the human-originated texts look more irregular, with negative lag-1 autocorrelation in entropy, than those of the human-mimicking AI text, but this is not statistically supported with our sample size and is not consistent across both AI models. We hope that this study encourages larger experiments with multiple authors, topics and evaluator models.

Declarations

- Funding: The author confirms that this research did not receive any external funding.
- Competing interests: The author declares no competing interests.
- Data availability: Not shared publicly, available upon request.
- Code availability: Not shared publicly, available upon request.
- Author contribution: This study was conducted by the author solely.

Appendix A Sentence-Level PPL Table

Table A1: Sentence-Level Perplexity Scores Across Writing Conditions Using Qwen3

	Human-Originated Writing			AI-Generated Writing	
Sentence	**Human**	**Grammar**	**Refined**	**GLM Mimicking**	**GLM Flash**
1	89.9	89.9	40.5	79.1	93.6
2	45.7	30.4	15.2	27.6	49.2
3	54.4	73.5	61.4	27.1	30.3
4	44.2	36.6	23.1	32.9	28.7
5	37.0	35.5	12.3	23.2	35.5
6	9.0	10.1	35.4	32.5	77.9
7	23.3	22.8	12.2	20.3	9.9
8	38.5	33.2	27.8	24.4	48.7
9	23.9	23.5	10.3	25.8	13.0
10	34.6	26.3	9.8	19.7	37.4
11	59.5	52.7	17.4	10.9	27.0
12	28.7	22.1	6.0	14.7	18.7
13	26.0	27.5	5.0	11.8	17.9

Note. Higher PPL values indicate greater model uncertainty in predicting the corresponding text under the selected language model.

Appendix B Sentence-Level Entropy Table

Table B2: Sentence-Level Entropy (bits) Across Writing Conditions Using Qwen3

	Human-Originated Writing			AI-Generated Writing	
Sentence	**Human**	**Grammar**	**Refined**	**GLM Mimicking**	**GLM Flash**
1	3.56	3.56	2.90	3.38	2.72
2	2.23	2.35	2.12	2.36	2.26
3	3.10	3.23	1.79	1.70	2.05
4	3.02	3.20	2.18	2.10	2.42
5	2.98	3.05	2.08	2.45	2.73
6	3.03	3.19	2.91	2.80	3.24
7	2.57	2.47	2.22	2.85	2.72
8	2.76	2.73	2.30	3.18	3.23
9	3.01	3.11	2.08	2.97	2.51
10	2.87	3.35	2.25	2.64	3.03
11	3.41	3.50	2.09	2.50	2.97
12	3.32	3.12	2.43	2.23	3.28
13	2.81	2.72	1.89	2.54	2.34

Appendix C Sentence-Level Entropy Net Change Table

Table C3: Sentence-to-Sentence Net Entropy Changes Across Response Categories

Transition	**Human**	**Grammar-Corr.**	**Human-Refined**	**GLM 5.3 Mimic**	**GLM-5.3-Flash**
Sent 1 → 2	−1.33	−1.21	−0.78	−1.02	−0.46
Sent 2 → 3	0.87	0.88	−0.33	−0.66	−0.21
Sent 3 → 4	−0.08	−0.03	0.39	0.40	0.37
Sent 4 → 5	−0.04	−0.15	−0.10	0.35	0.31
Sent 5 → 6	0.05	0.14	0.83	0.35	0.51
Sent 6 → 7	−0.46	−0.72	−0.69	0.05	−0.52
Sent 7 → 8	0.19	0.26	0.08	0.33	0.51
Sent 8 → 9	0.25	0.38	−0.22	−0.21	−0.72
Sent 9 → 10	−0.14	0.24	0.17	−0.33	0.52
Sent 10 → 11	0.54	0.15	−0.16	−0.14	−0.06
Sent 11 → 12	−0.09	−0.38	0.34	−0.27	0.31
Sent 12 → 13	−0.51	−0.40	−0.54	0.31	−0.94

Appendix D Prompts Used for the AI-Generated Responses

The prompts are reproduced as they were given. The writing sample inside the first prompt (the grammar-corrected human essay) is omitted here and replaced by a placeholder.

GLM 5.3 Mimicking.

```
I want you to write in my voice, based on the writing sample below. Your goal is for the result to
     read like I wrote it myself, not like an AI.

First, study the sample and note (privately, no need to list them back to me):
- Sentence length and rhythm (short and punchy, long and flowing, or a mix)
- Vocabulary level and favorite words or phrases
- Tone (casual, formal, dry, warm, sarcastic, etc.)
- How I open and close pieces
- Punctuation habits (commas, dashes, ellipses, exclamation marks, parentheses)
- Quirks: contractions, fragments, slang, rhetorical questions, humor, typos or informality
- Paragraph length and structure
- How much I explain versus assume the reader knows

Then write the piece below, following these rules:
1. Match my style closely. Mimic the patterns, don't copy sentences from the sample.
2. Avoid typical AI tells: "delve," "tapestry," "in today's fast-paced world," "it's important to
     note," "furthermore," "in conclusion," tidy rule-of-three lists, overly balanced "not only...
     but also" constructions, and neat summary endings.
3. Don't over-polish. Real writing has uneven rhythm, the occasional tangent, and opinions stated
     plainly.
4. Use concrete details and specific examples instead of generic statements.
5. Don't use headings, bullet points, or bold unless my sample does.

--- MY WRITING SAMPLE ---
[grammar-corrected human essay, 13 sentences, omitted]
--- END SAMPLE ---

--- WHAT I WANT YOU TO WRITE ---
Topic/purpose: Some people believe that university education should be free for everyone, regardless
      of their financial background. To what extent do you agree or disagree?
Audience: IELTS Examiner
Key points: do not just copy-paste the things of my sample; you can adopt some but not fully
Maximum Sentence : 13 and no less or more is accepted.
Word Limits : 250 - 350
--- END ---

After writing, don't explain your choices. Just give me the finished text.
```

GLM-5.3-Flash.

```
--- WHAT I WANT YOU TO WRITE ---
Topic/purpose: Some people believe that university education should be free for everyone, regardless
      of their financial background. To what extent do you agree or disagree?
Audience: IELTS Examiner
Key points: do not just copy-paste the things of my sample memory or any available online source;
     force your own knowledge to write
Maximum Sentence: exactly 13, and no less or more is accepted.
Word Limits: 250 - 350
--- END ---

After writing, don't explain your choices. Just give me the finished text.
```

References

[1] Jelinek, F., Mercer, R.L., Bahl, L.R., Baker, J.K.: Perplexity—a measure of the difficulty of speech recognition tasks. The Journal of the Acoustical Society of America **62**(S1), 63 (1977)

[2] Gutiérrez Megías, A.J., Ureña-López, L.A., Martínez Cámara, E.: The influence of the perplexity score in the detection of machine-generated texts. In: Mitkov, R., Ezzini, S., Ranasinghe, T., Ezeani, I., Khallaf, N., Acarturk, C., Bradbury, M., El-Haj, M., Rayson, P. (eds.) Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, pp. 80–85. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, Lancaster, UK (2024). https://aclanthology.org/2024.nlpaics-1.10/

[3] Li, Y., Li, Q., Cui, L., Bi, W., Wang, Z., Wang, L., Yang, L., Shi, S., Zhang, Y.: MAGE: Machine-generated text detection in the wild. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 36–53. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.3 . https://aclanthology.org/2024.acl-long.3/

[4] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., Zou, J.: Gpt detectors are biased against non-native english writers. Patterns **4**(7), 100779–100779 (2023) https://doi.org/10.1016/j.patter.2023.100779

[5] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., Waddington, L.: Testing of detection tools for AI-generated text. International Journal for Educational Integrity **19**(1), 26 (2023) https://doi.org/10.1007/s40979-023-00146-z

[6] Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**(3), 379–423 (1948)

[7] Wang, X., Rushing, B., Chen, Z., Zhe, Z.: Local Entropy Fluctuations for Black-Box LLM-Generated Text Detection. Accessed 2026-11-04 (2026). https://arxiv.org/abs/2610.33767v2

[8] Gehrmann, S., Strobelt, H., Rush, A.: GLTR: Statistical detection and visualization of generated text. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 111–116. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-3019

[9] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C.: DetectGPT: Zero-shot machine-generated text detection using probability curvature. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings

of Machine Learning Research, vol. 202, pp. 24950–24962 (2023)

[10] Bao, G., Zhao, Y., Teng, Z., Yang, L., Zhang, Y.: Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In: International Conference on Learning Representations (2024)

[11] Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., Goldstein, T.: Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235 (2024)

[12] Ippolito, D., Duckworth, D., Callison-Burch, C., Eck, D.: Automatic detection of generated text is easiest when humans are fooled. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1808–1822. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.164

[13] Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=rygGQyrFvH

[14] Krishna, K., Song, Y., Karpinska, M., Wieting, J., Iyyer, M.: Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In: Advances in Neural Information Processing Systems, vol. 36 (2023)

[15] Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S.: Can AI-Generated Text be Reliably Detected? (2023). https://arxiv.org/abs/2303.11156

[16] Genzel, D., Charniak, E.: Entropy rate constancy in text. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 199–206 (2002)

[17] Yang, A., et al.: Qwen3 Technical Report (2025). https://arxiv.org/abs/2505.09388

[18] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control, 5th edn. John Wiley & Sons, Hoboken, NJ (2015)