Is Human-Like Text Liked by Humans? Multilingual Human Detection and Preference Against AI

Anonymous ACL submission

Abstract

Prior studies have shown that distinguishing text generated by large language models (LLMs) from human-written one is highly challenging, and often no better than random guessing. To verify the generalizability of this finding across languages and domains, we perform an extensive case study to identify the upper bound of human detection accuracy. Across 16 datasets covering 9 languages and 9 domains, 19 annotators achieved an average detection accuracy of 87.6%, thus challenging previous conclusions. We find that major gaps between human and machine text lie in concreteness, cultural nuances, and diversity. Prompting by explicitly explaining the distinctions in the prompts can partially bridge the gaps in over 50% of the cases. However, we also find that humans do not always prefer human-written text, particularly when they cannot clearly identify its source.

1 Introduction

Recent technological developments have advanced the generative artificial intelligence (AI) models, such as GPT-*, Gemini, Claude, and Llama (OpenAI, 2023; Team et al., 2023; Anthropic, 2024; Dubey et al., 2024), thus aggressively blurring the lines between human and AI capabilities. How often can state-of-the-art (SOTA) LLMs fool human evaluators, i.e., passing the Turing test? Can human evaluators correctly predict the origins of encountered content (human vs. machine), safeguarding against the potential misuse of LLMs?

Several studies have explored whether humans can distinguish between content generated by a human vs. a machine. Unsurprisingly, the findings depend in part on the capability of generative models. Early studies probing the outputs of less advanced AI models found that human evaluators could indeed detect the difference (van der Lee et al., 2019). However, in studies using LLMs like GPT-3.5-turbo, human evaluators are frequently close to random chance (Guo et al., 2023; Hitsuwari et al., 2023; Dugan et al., 2023; Chein et al., 2024; Wang et al., 2024a), particularly for laymen who are not frequent users of LLMs.

While covering different domains, e.g., academic paper abstracts, Wikipedia paragraphs, question-answering responses (Wang et al., 2024b,a), news and review comments (Chein et al., 2024), most studies focused on GPT-3.5-turbo and English texts (see Table 1). Generally, fewer than 300 examples were evaluated by experts (LLM researchers or frequent users) in such studies.

With the advancement of newer LLMs such as GPT-4o, Claude-3.5-Sonnet, and Llama4, we ask whether these findings generalize to them, to other languages, and to native expert annotators? What are the **upper bounds of human detection** performance across languages, domains and new LLMs?

To answer these questions, we perform a comprehensive case study over 16 datasets spanning 9 languages, 9 domains, and 11 SOTA LLMs. We specifically focus on the following four research questions: (*i*) *How well can human annotators distinguish human-written text from SOTA LLMs generations?* (*ii*) *What are the notably detectable linguistic qualities in human- and AI-authored texts that shape decisions about text origin?* (*iii*) *Can a prompting strategy fill the gap between human and machine?* and (*iv*) *Do humans always prefer human-authored text?*

Our comprehensive analysis exposes the major gaps between current LLMs and human expectations, suggesting possible directions for future LLM development. Our human detection evaluation challenges previous findings in terms of random-guess detection capability for humans. Moreover, our human preference collection can serve as a potential useful source for multilingual LLM alignment. Our contributions are as follows:

- Our extensive case study shows that expert annotators achieve detection accuracy of **87.6%**

on average over ∼9K examples, demonstrating that humans can identify LLM outputs and safeguard against LLM misuse. We identified a spectrum of language- and domain-specific distinguishable features that distinguish human-written from machine-generated text, and we summarized five generalized distinctions across languages, which clearly expose the limitations of current LLM outputs and indicate directions for improvements.

- Prompting by explicitly explaining the gap patterns between human text and LLM generations in instructions can either fully or partially address issues for over 50% of the cases. However, cultural nuances, diversity in length, structure, and sentiment remain challenging.
- Humans do not always like human-written text: preferences are evenly split between human vs. machine, with a strong tendency to favor machine-generated text (MGT) when annotators can not clearly identify the origin of a given text.
- We release 17k original and 32k improved MGTs, labels of 13.5k human detection instances, 5k preferences and 1.6k fill-in-the-gap survey, and metadata for 19 annotators to facilitate future work on human vs. MGT detection, preference, and their relationship with individual characteristics.

2 Case Study

Previous studies presented the difficulty humans face in distinguishing SOTA LLM-generated content from human-written text, often resulting in a random guess (see more in Appendix A). However, most evaluations focused on English and generations by GPT-3.5-turbo, leaving the detectability of MGT in other languages and LLMs uncertain.

To verify whether this observation can generalize to other languages and more advanced LLMs, we collected LLM generations based on 16 datasets across 9 domains and 9 languages. 19 native speakers who are either LLM researchers or practitioners performed human evaluations, investigating (*i*) whether expert humans can correctly discern human vs. AI outputs, and (*ii*) whether humans prefer fellow human answers or LLM responses.

2.1 Task and Dataset

MGT detection The goal is to identify whether the text was written by a human or generated by models given a single text, or to recognize which text is written by a human given a pair of texts: one human-written and one machine-generated.

In data collection, we focus on datasets from common domains such as community QA, news, tweets, and government reports, alongside domains requiring high-integrity LLM applications, including educational and academic contexts, such as accurate knowledge verification in Wikipedia-style texts, and identifying the authorship of student essays and peer reviews.

For a given language and given a dataset, we sampled 300-600 human texts and then generated corresponding machine text using two SOTA LLMs: a multilingual model (e.g., GPT or Claude series) and a language-specific model (e.g., ChatGLM or Qwen for Chinese and AceGPT for Arabic), to analyze the impact of different LLMs on detection performance, particularly for non-English languages. Table 4 shows the statistical information of the 16 datasets, and Appendix C elaborates the generation prompts and collection details.

2.2 Human Detection Setups

Annotation Settings To mimic real-world MGT detection scenarios, we set up four human evaluation settings. Given human-written text and machine-generated text, representing by `hwt` and `mgt` respectively, human annotators are asked to identify which text was written by a human. Note that `mgt` can be generated by multiple different LLMs, referring to as $\texttt{mgt}_\texttt{i}$, where $\texttt{i} \in \mathbb{N}$.

As shown in Table 2, according to the input and the output options, we categorize detection settings as I. *pair-binary*, II. *pair-four-class*, III. *single-binary*, and IV. *triplet-three-class*. For a single text input, either `hwt` or `mgt`, the goal is to recognize whether the text was written by a human, by answering just Yes or No. This is suitable for the scenario where for the human text there is no necessarily a corresponding machine-generated text, and thus they can be collected from different sources and for different topics, such as Arabic tweets. Given a pair of texts (text1, text2), a binary output is easier than the four-class detection. The *pair-binary* setting asks that either text1 or text2 is `hwt`, and the other one is `mgt`, while the *pair-four-class* setting has no restrictions: each of text1 and text2 can be `hwt` or `mgt`, regardless of the label of the other text. Sometimes, we want to compare human text to generations from different LLMs,

Study/ Paper	MGT Detection Task Formulation	Lang-uage	Generative Models	MGT Domain	#Sample	Human Annotators Where and Size	Native	Layman	Avg.Acc/ Range
			Before OpenAI released GPT-3.5-turbo in November 2022						
Garbacea et al. (2019)	2-class	English	12 sequence models before 2019	online product reviews	3,600	AMT-900	–	✓	66.61%
Ippolito et al. (2020)	2-class	English	GPT2-large	web text	150	AMT, uni students	✓	✗	71.4%
Clark et al. (2021)	2-class	English	GPT2, GPT3	stories, news articles, recipes	3,900	AMT-780	✓	✓	50-58%
Shamardina et al. (2022, 2025)	2-class	Russian	14 decoder-only, encoder-decoder models	Wikipedia, news, social media, RNC, diaries, strategic documents	2,500	Toloka	✓	✓	66.66%
			After OpenAI GPT-3.5-turbo						
Guo et al. (2023)	2-class	English	GPT-3.5	Wikipedia, QA	150	8 experts + 9 layman	✗	✓	48-90%
Guo et al. (2023)	2-class	Chinese	GPT-3.5	Wikipedia, QA	210	8 experts + 9 layman	✓	✓	54-93%
Chein et al. (2024)	2-class	English	GPT-3.5-turbo	news	96	203 Prolific	✓	✓	57%
Chein et al. (2024)	2-class	English	GPT-3.5-turbo	social media comments	96	203 Prolific	✓	✓	78%
Wang et al. (2024b)	2-class	English	GPT-3.5-turbo	Reddit and arXiv abstract	100	6 experts	✗	✗	41-94%
Dugan et al. (2023)	boundary detection	English	GPT2 (-XL), CTRL	news, stories, recipes, speeches	7,257	RoFT game player	–	✓	23.4%
Wang et al. (2024a)	4-class	English	davinci-text-003, GPT-3.5-turbo, Cohere, Dolly-v2, BLOOMz, GPT-4	Wikipedia, WikiHow, Reddit, arXiv abstract, peer review	140	4 experts	✗	✗	10.4-27.4%

Table 1: *Experts* refers to individuals who are frequent users of LLMs, *laymen* are people who have never heard of or used LLMs or have rarely used them, *AMT* is Amazon Mechanical Turk. *RNC* is the Russian national corpus.

Setting ID	Input	Task Description	Output Options	Applicable Scenarios
I. Pair-Binary	(hwt, mgt) **or** (mgt, hwt)	Given a pair of (text1, text2), identify which one is human-written? Either text1 or text2 must be hwt, and another is mgt randomly sampled from mgt_i.	**A.** text1; **B.** text2	parallel data is available.
II. Pair-Four-Class	(hwt, mgt) **or** (mgt, hwt) **or** (hwt, hwt) **or** (mgt, mgt)	Given a pair of (text1, text2), identify which one is human-written? Both text1 and text2 can be hwt, and can be mgt.	**A.** text1; **B.** text2; **C.** none of them; **D.** both	parallel data is not necessary.
III. Single-Binary	hwt or mgt	Given a piece of text, identify whether it is written by human?	**A.** Yes, human; **B.** No, machine	parallel data is not necessary.
IV. Triplet-Three-Class	(hwt, mgt_1, mgt_2)	Given a set of texts (text1, text2, text3), identify which one is human-written? One of the text1, text2 and text3 must be hwt, and others are mgt randomly sampled from mgt_i.	**A.** text1; **B.** text2; **C.** text3	parallel human and multiple LLM generations are collected to make comparisons.

Table 2: The four human detection settings: the setting name refers to input-output options, pair/binary means the input is a pair of texts and the goal is to predict binary labels, and whether the text is human-written (Yes or No).

in which case, we apply IV, which we limit to a three-class detection: human vs. LLM_1 vs. LLM_2.

Overall, I and IV are applicable in scenarios where a human text has a corresponding machine-generated counterpart. II and III can be utilized when the human- and machine-generated texts are non-corresponding. If annotators are exposed to some human- and machine-written texts prior to the detection task, we refer to it as a few-shot setting; otherwise, it is considered zero-shot.

Annotator Background We highlight that this study aims to explore the upper bound of human detection capability, rather than represent the full population distribution. Therefore, we did not recruit lay annotators unfamiliar with LLMs. Instead, we conducted in-house annotation with BSc., MSc., and PhD. students, as well as postdoctoral researchers, all with NLP expertise and prior experience with LLM-generated texts. Annotators were native speakers of the respective languages and completed annotations independently (see Appendix A.3 for details).

While using expert annotators may limit the generalizability of results to lay users, our goal is to establish expert performance ceilings. If trained experts struggle to detect machine-generated text, it is reasonable to expect lower performance from the general public. Lay annotator studies are important for practical applications but do not substitute for upper-bound evaluation. Future work will include lay annotators to capture a broader distribution of human detection capabilities.

3 Human Detection

We performed an extensive case study on 9k instances across 9 languages to verify how difficult it is for native speakers to detect AI outputs in everyday domains. Table 3 demonstrates that human detection accuracy can reach 87.6% on average. This reveals that this is not particularly difficult for native human experts, contrary to what previous studies have reported. Below, we zoom into the impact of various factors.

3.1 Language

Human detection accuracy exceeds 87.6% for Chinese, English, Arabic, Italian and Russian, while it falls below this level for Vietnamese, Kazakh, Hindi, and Japanese. This discrepancy is largely due to the challenge of Wikipedia text.

3.2 Domain

Wikipedia is widely used as training data for LLMs, particularly for low-resource languages, due to the scarcity of alternative datasets. Consequently, models often memorize Wikipedia content, leading to generated text that closely resembles human-written Wikipedia. Arabic tweets also present challenges for detection due to their short length and

Language	Source/Model	#Example	#Annotator	Anno Setup	Shot	Avg. Acc
Arabic	Dialect Tweet	900	1	III. Single-binary	Zero	50.1
	EASC Summary	100	1	I. Pair-binary	Zero	82.0
	Youm7 News	1,000	1	I. Pair-binary	Zero	92.7
	SANAD News	100	1	I. Pair-binary	Zero	100.0
Chinese	Zhihu-QA (GPT-4o)	428	5	I. Pair-binary	Zero	99.6
	Zhihu-QA (GPT-4o)	160	1	I. Pair-binary	Few	100.0
	Zhihu-QA (Qwen-turbo)	588	2	I. Pair-binary	Zero	98.0
	Student essay	102	1	I. Pair-binary	Few	98.0
	Student essay	600	3	II. Pair-four-class	Zero	97.0
	Government Report	500	1	IV. Triplet-three-class	Few	97.2
English	Peersum	400	1	I. Pair-binary	Few	99.8
Hindi	News	600	1	I. Pair-binary	Few	85.2
Italian	DICE News (Anita)	300	1	I. Pair-binary	Few	88.0
	DICE News (Llama3-405B)	300	1	I. Pair-binary	Few	99.7
	DICE News (GPT-4o)	300	1	I. Pair-binary	Few	100.0
Japanese	News	300	2	I. Pair-binary	Zero	86.4
Kazakh	Wikipedia	300	2	I. Pair-binary	Zero	79.7
Russian	News	300	1	I. Pair-binary	Few	100.0
	Academic summary	300	1	III. Single-binary	Few	80.0
Vietnamese	Wikipedia	600	1	I. Pair-binary	Zero	50.7
	News	600	1	I. Pair-binary	Zero	80.3
Total	–	8,778	30			**87.6**

Table 3: Human annotator detection accuracy over 16 datasets and 9 languages: we have a total of 30 annotation settings and 19 unique human annotators. The simple average accuracy of the human expert guesses is 87.6%.

limited context, along with summaries, e.g., for Arabic and Russian summaries, the expert detection accuracy is about 80%. This conversely highlights the ability of language models to generate high-quality human-like text in the domains of Wikipedia, tweets, and summaries. In contrast, substantial differences between machine-generated and human-written text persist in news articles, QA, student essays, and peer reviews, making them much easier to recognize for human experts.

3.3 Generator

The impact of generators is associated with languages and domains. While there are minimal differences for Chinese, where accuracy is 100% vs. 98% for GPT-4o vs. Qwen-turbo, there are sizable differences for Italian and Arabic. Based on Italian DICE News, the same annotator detected generations by Anita (an Italian fine-tuned Llama3-8B), Llama3-405B, and GPT-4o, achieving accuracy of 88%, 99%, and 100%, respectively. Similarly, for Arabic tweets, GPT-4o's outputs are more similar to human text and thus more difficult to detect compared to those by Qwen2, as shown in Table 6.

3.4 Annotation Setting

We conducted the majority of our annotations under setting I. *pair-binary*: given a pair (`hwt`, `mgt`), it asks to identify which of the two texts is human-written. We assumed that more complex settings would be more challenging. For instance, II. *pair-four-class* should be harder, as each of the texts could be human- or machine-generated, independently of the other. Yet, for Chinese student essays, the performance for II does not degrade compared to I. Similarly, for government reports in IV. *triplet-three-class*, where the annotators have to select the human-written text among three candidates, there was no degradation compared to I.

However, III. *single-binary* proves to be more difficult than I. *pair-binary* for both Arabic and Russian. While domain differences do exist, e.g., tweets vs. summary vs. news in Arabic, and news vs. summary for Russian, the substantial performance gap (>20%) can still be partially attributed to the annotation settings. Overall, comparing the results for I. vs. III., it is easier to distinguish machine-generated content when given a comparison pair, rather than for single answer. Yet, introducing text triplets or increasing the number of machine-generated or human-written texts had minimal impact on detection performance. Moreover, using few shots before detection boosted the confidence of the annotators, resulting in higher accuracy compared to zero-shot. For datasets with high accuracy, before seeing labeled samples, the anno-

tators found the distinction to be obvious. After seeing a few examples, the annotator was extremely confident in distinguishing human vs. machine text based on indicative features of MGT.

3.5 Expert Annotators

For the same language and dataset, individual annotator ability influences detection accuracy but not significantly. For instance, in Chinese Zhihu-QA (GPT-4o vs. human), five annotators achieved accuracies of 99%, 99%, 100%, 100%, and 100%. Similarly, for Zhihu-QA (Qwen-turbo vs. human), two annotators obtained 99% and 97%. In student essays (II. pair-four-class), three annotators recorded accuracies of 96%, 96%, and 99%. This may also result from the bias that all annotators are native speakers and expert-level LLM practitioners or researchers. Differences between individuals are minor in their cognitive abilities, language proficiency and domain knowledge.

3.6 Distinguishable Factors

We summarize five remarkable distinguishable signals between machine-generated vs. human-written text across the 16 datasets and the 9 languages. Appendix D provides detailed distinctions between `hwt` and `mgt` for each language and dataset, where valuable expert insights into the limitations and gaps of current LLM outputs compared to human-written text can inform future model improvements.

- **Human text is more informative and concrete.** Human-written text contains concrete numbers, specific names of people or institutions, exact places or dates, URLs, and other references, while machine-generated text tends to provide generic information, with little detail to support its statements.
- **Machine-generated text lacks regional, cultural, and religious nuances.** For languages such as Arabic, Japanese, Hindi, Kazakh, and Chinese, human texts reflect the cultural and the religious nuances of the language, which is not true for machine-generated text.
- **Human-written text varies substantially in terms of length, structure, style, and sentiment.** Human texts show diversity and inconsistency with large deviations in length, structure, style and emotions, while machine-generated texts tend to use a formulaic structure and neutral/positive emotion. This can be partially attributed to LLMs rigorously following instructions, and thus losing on flexibility.
- **Machine-generated text has formatting.** MGTs are generally well-segmented with bullet points for better readability, while human-written texts are typically just large block of plain text, which may be due to human text collection and conversion. Moreover, machine-generated texts often use Markdown style, e.g., ** and ###, while human-written texts have typos, grammatical errors, hashtags, and other social media elements.
- **Machine-generated text shows a mixture of other languages.** Non-English language responses often contain some English parts, and sometimes mix other languages, e.g., Japanese responses contain Chinese or Korean, particularly for less capable models. This is very rare in human text.

4 Can Prompting Fill in the Gap?

Given that LLMs can closely follow instructions and their outputs are heavily influenced by the system and the user prompts, we investigated whether explicitly instructing them to mimic human style can help narrow the gap. Responding to the distinguishable factors summarized in Appendix D for each dataset, we asked the human annotators to craft new prompts, aiming to improve the generations and to reduce the gap between human-written and LLM-generated texts. This involved trying instructions that (*i*) incorporate specific details and references, (*ii*) avoid formulaic structures and formats, e.g., bullet points and Markdown, and (*iii*) generate texts of varying length, structure, and sentiment. Table 8 presents both the original and the improved prompts for all datasets.

Measurements: We re-generated the machine-generated parts of the text pairs, using the same models with improved prompts, and then sampled 200–600 examples from each dataset to assess whether and to what extent, the prompting strategy narrowed the gap between human-written and machine-generated texts. We used two approaches: (*i*) intuition-oriented *fill-the-gap survey*, where the original annotators evaluated whether the newly-generated text bridged the gap (Yes, No, or Partially), and (*ii*) *downstream detection*, where we compared the detection accuracy before and after applying the improved prompts. A decline in detection accuracy indicated a reduced distinction between human and machine text, making the differentiation more difficult, and further revealing

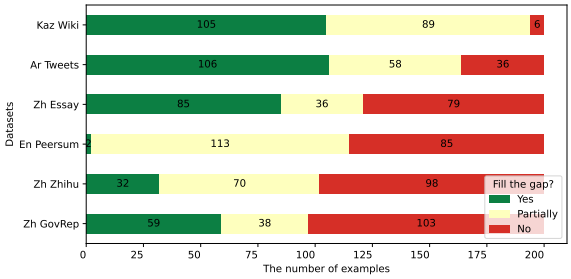

Figure 1: Evaluating whether the new generations fill in the gap: Yes, Partially, or No.

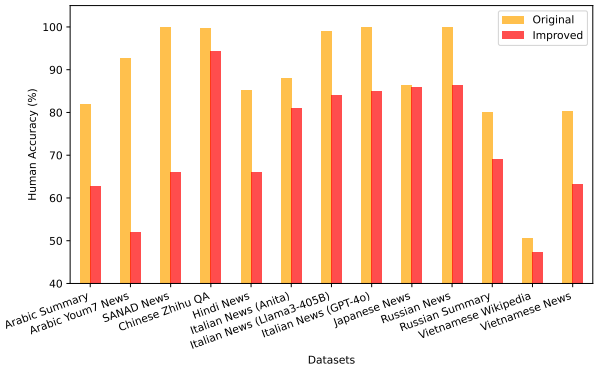

Figure 2: Human detection accuracy dropped from 87.6 to 72.5 for the original vs. the improved generations.

that prompting was effective. While the survey is inherently slightly subjective, it remains valuable, as human intuition plays a crucial role in evaluating the perceived quality of model outputs. By combining this intuition-driven assessment with quantitative detection results, we aim to draw more robust and convincing conclusions.

Fill-the-gap Survey The original annotators who conducted the detection on earlier generations were asked to evaluate whether the new prompts addressed the gaps for each example: *Yes*, *No*, and *Partially*. The distributions across the six datasets in Figure 1 shows that in about 50% of the cases, the prompt adjustments were effective to either fully or partially mitigate the gaps. Large improvements were observed for Kazakh Wikipedia and Arabic tweets. For the former, the revised prompt reduced repetitive sentence patterns (more diverse), but the formulaic expressions were not entirely eliminated. New outputs also included more concrete information, such as dates and names, while the inclusion of culturally-nuanced details remained challenging. The newly-generated Arabic tweets could touch on relatable human topics and express genuine emotions tied to daily experiences; however, the frequently added irrelevant hashtags at the end of the tweets made them easily identifiable as machine-generated. Moreover, the tweets often leaned on an overly optimistic tone even when negative experiences were mentioned.

The annotators for English peer reviews noted that, despite the prompt adjustments, the models remained highly formulaic in their outputs, the length of the reviews remained relatively uniform, and the overall structure lacked variance. This may be due to the inherent nature of the peer review domain, while human reviews exhibited more variability in both length and structure, feeling more organic. Similar issues remained for Chinese student essays (formulaic structure by using "firstly, then, moreover, finally" persisted) and government reports (certain repetitive phrases). See more details in Appendix D.2.

Overall, adjusting the prompts did fill some gaps, but cultural nuances, diversity of length, structure and phrases, and sentiment adaption to scenarios remained challenging. See more in Appendix E. With the intuitive survey results, we further conducted a second round of MGT detection on the new generations for objective results, involving both human evaluation and automated detection.

MGT Detection on Improved Text We performed human detection on 13 datasets under the same annotation setting as described in Table 3, with the exact same annotators. We observed sizable accuracy declines across all datasets in Figure 2, with average accuracy dropping to 72.5%. This implies that the improved machine-generated text became more similar to the human-written text, making it harder to discern and thus resulting in lower detection accuracy.

We further analyzed the impact of the improved generations on automatic detection accuracies. We collected a total of 17,017 texts using the original prompts and 32,487 texts using the improved prompts (detailed statistical distribution in Table 9). We reproduced 26 MGT detection approaches presented in the COLING 2025 GenAI shared task (Wang et al., 2025) and evaluated them on the collected MGTs. As shown in Figure 6, 19 methods exhibited lower accuracy on the newly generated texts, indicating that the texts produced using the improved prompts are more challenging to distinguish compared to those generated with the original prompts. This suggests that prompt-

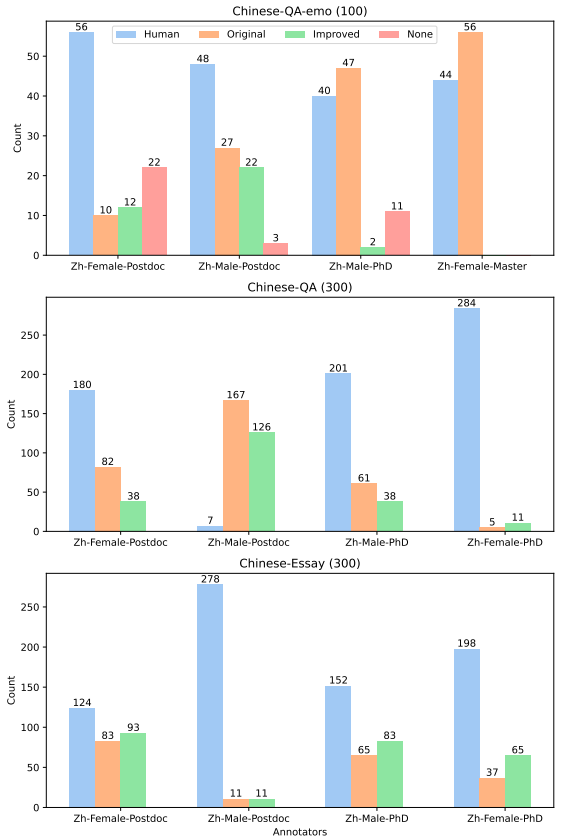

Figure 3: Human preferences for three Chinese datasets (five annotators): QA-emo is an emotion-rich question subset of Zhihu-QA with 100 examples.

ing strategies can help bridge some gaps between machine-generated and human-written text when explicitly designed to mimic human writing style.

5 Human-Like or Liked-by-Human?

We used the prompting strategy to bridge the gap between human and machine text, aiming to make machine outputs more human-like. However, do humans favor human-like text? Below, we examine human preferences among four options: (*i*) human-written text, (*ii*) machine-generated text using the original prompt, (*iii*) machine-generated text using the improved prompt, or (*vi*) none of the above.

Preference Labeling Setup: To evaluate the impact of human detection accuracy on their preferences, we selected datasets with high accuracy (Chinese QA and essays, Russian news), medium accuracy (Russian and Arabic summaries), and low accuracy (Arabic tweets). For Chinese, we annotated Zhihu-QA and student essays (300 examples for each), along with 100 responses particularly for Zhihu questions, where emotional and empathetic comforts are highlighted. Five unique annotators participated, identified by *nationality-gender-degree*. For example, *Zh-Male-PhD* refers to a Chinese male PhD student. We labeled Russian with three annotators and two annotators for Arabic.

Do People Always Prefer Human Text? The answer is *No*. Analyzing the preferences of ten annotators across six datasets in Figure 3 and 4, human text was preferred in about half of the cases. Notably, for Russian and Arabic, annotators tended to favor machine text. This is particularly evident for Russian summaries using the improved prompts (green bars) and Arabic summaries using the original prompts (orange bars).

Annotators prefer human-written text when they can clearly identify its source, otherwise they tend to favor machine text. For Chinese Zhihu QA and student essays, where detection accuracy approaches 100%, human-written text is generally preferred though there are exceptions. *Zh-Male-postdoc* exhibits a unique preference distribution that deviates from the rest. *Ru-Male-PhD* achieved 100% accuracy on Russian news and favored human-written text, while the other two annotators, with only medium accuracy, preferred machine-generated text. For datasets with medium or low detection accuracy (Russian and Arabic summaries, Arabic tweets), machine text was preferred in $\geq \frac{2}{3}$ of the cases. These suggest that when annotators can clearly identify human-written content, social or psychological bias toward peer-authored text emerges, in contrast, when uncertain, their intuitive preferences tend to favor machine outputs.

Are Humans Good Emotional Support? *Not necessary.* Interestingly, for emotion-rich questions (QA-emo), where human responses are expected to be more empathetic and be preferred, three out of four annotators actually preferred the machine-generated responses. The remaining annotator disliked both human and machine answers in 22% of the cases. The annotator feedback suggested that this was influenced by the presence of mean-spirited responses from Zhihu users, where some human answers expressed personal biases and lacked empathy (see Appendix G.1).

Why are Human-Written Essays Favored? Human-written essays tend to exhibit greater coherence and sincerity. In contrast, machine-generated essays often lack cohesion, featuring abrupt transitions and sometimes repeating the title in an at-

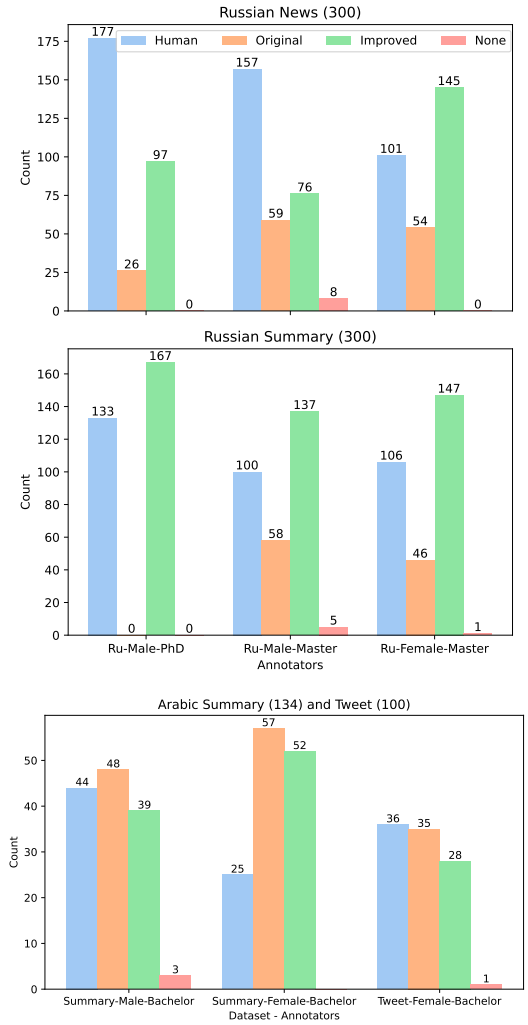

Figure 4: Human preferences for two Russian (three annotators) and two Arabic datasets (two annotators).

tempt to simulate continuity. While LLMs can generate related concepts, they struggle to construct a compelling narrative or to seamlessly connect multiple stories under a unified theme. As a result, their outputs often feel shallow—-relying on abstract concepts and rhetorical flourishes without offering substantive insight. The text tends to be verbose yet superficial, giving an impression of grandiosity without meaning. Moreover, LLM-generated essays often adopt an instructive tone, resembling answers to questions rather than peer-level discourse.

From a literary perspective, human-written articles often charm with playfulness, sincerity, or novel insights, using precise language to evoke emotion and poetic depth. In contrast, a piece generated by an LLM lacks this literary nuance, leaving little room for contemplation or lingering thoughts. The expression is immediately clear, requiring no further reflection, resembling more of a lecture than an engaging discourse.

Human-Like or Liked-by-Human? Human texts are not always preferred. This inspires us to reflect the ultimate goal of building LLMs that are human-like vs. liked-by-human. The goal of being human-like has a single target, i.e., mimicking human behavior, while to be liked-by-humans involves optimization towards billions of local optima, each shaped by individual preferences. Current language models establish a foundation by learning from human data towards being human-like. As they get more advanced, they can further adapt by incorporating personal data, thus transitioning from merely imitating human behavior to aligning with individual user preferences, and moving from human-like to liked-by-human. Our preference annotations can serve as a valuable resource for guiding models to match individual preferences in multilingual contexts. Also, the data can facilitate investigation of the relationship between annotator characteristics (e.g., MBTI personality, gender, and age) and preferences.

6 Conclusion and Future Work

We conducted a comprehensive study to investigate the upper bound capability of humans to detect text generated by SOTA LLMs. Based on 16 datasets spanning 9 languages, 9 domains, and 11 LLMs, 19 expert annotators' accuracy ranged in 50-100%, with an average of 87.6%, revealing that it is not that challenging for experts to identify MGTs.

We found that the major gaps between human-written and machine-generated text lie in concreteness, cultural nuances, diversity in length, structure, style, and sentiment, formatting and mixture of languages in the generations. Prompting by explicitly indicating the distinctions could partially bridge this gap, but cultural nuances and diversity remained a challenge. Yet, humans favored machine-generated text in over half of the cases, particularly when they could not recognize which text was written by a human.

In future work, we plan to explore human detection accuracy and preference across diverse population groups, especially their relationship with individual characteristics. How do an individual's philosophy, personality, and experience influence the detection ability and preferences.

Limitations

Annotator Diversity All annotators involved in the detection process are advanced NLP researchers, including Msc. and PhD. students, as well as postdocs, specializing in LLMs. No laymen participated in the annotation process, which ensures that the findings and the conclusions were relevant within the expert domain. It is possible that different results could be obtained if laypersons were involved in the detection tasks. However, given our aim to explore the upper bound of human capability in detecting machine-generated text, the current setting is appropriate. Future work could consider involving a broader range of participants to assess the average detection capability across different population groups.

Automatic Linguistic Analysis We did not include automated linguistic analysis using tools and techniques in this study. The reason lies in that our primary focus is on human-centered case study: how humans perceive quality of the output, identify distinguishable signals between human and machine, detection capabilities and preferences. There have been many studies exploring automatic MGT detection and performing linguistic analysis by automatic methods, while our aim is to understand human perception. Appendix D provides detailed linguistic differences (distinguishable signals) between human-written and machine-generated text which were manually summarized by expert human annotators. They are valuable insights to guide the future LLM improvement.

Our future work may explore more differences between human linguistics perceptions and insights derived from advanced computational analysis, across the three text types (human, original machine-generated, and improved machine-generated). Analysis would involve vocabulary features, part-of-speech tagging, dependency structures, sentiment analysis, language model perplexity, and other relevant metrics.

Statistical Significance In our study, each dataset includes at least 300 examples for annotation. For some datasets, 3-5 annotators labeled the same set of data, and the average values were calculated. However, a larger sample size would be beneficial to ensure more reliable results. In particular, in the survey exploring whether improved prompts bridge the gap between human-written and machine-generated text, substantial variability between individuals was observed. This subjectivity should be carefully considered in future work. Fortunately, three evaluation methods were used to assess whether prompting could mitigate this gap in our work, partially addressing this limitation by comparing human and automatic detection accuracy before and after the application of the improved prompts.

Ethical Statement

This section outlines potential ethical considerations related to our work.

Data Collection and Licenses A primary ethical consideration is the data license. We reused existing dataset for our research, such as HC3 M4GT-Bench, MAGE, RAID, OUTFOX and LLM-DetectAIve, which have been publicly released under clear licensing agreements. We adhere to the intended usage of these dataset licenses.

Security Implications The shared task datasets support the development of robust MGT detection systems, essential for addressing security and ethical concerns. These systems help prevent automated misinformation campaigns, protect against financial fraud, and ensure content integrity in critical domains like journalism, academia, and law. Beyond security, effective detection promotes digital literacy by fostering public awareness of LLM limitations and encouraging critical engagement with online content.

In multilingual contexts, MGT detection is particularly challenging due to linguistic and cultural nuances. Advanced systems must address these complexities in order to prevent disinformation, especially in less-resourced languages. Strengthening multilingual detection enhances the global trust in AI technologies and mitigates the security risks associated with their misuse.

Human Subject Considerations Our study engaged human evaluators in distinguishing between human-written and machine-generated texts, while expressing their preference. All annotators gave informed consent, were fully aware of the study's objectives, and had the right to withdraw at any time. Since this study was designed to be evaluated by experts, all annotators participating in it were advanced NLP researchers (MSc and PhD students, as well as postdocs, specializing in LLMs). While this ensures a rigorous assessment of the detection capabilities within an expert domain, we

acknowledge that findings may not generalize to laypeople. Future research should include a more diverse range of participants to evaluate human detection performance across various demographic and professional backgrounds.

Transparency and Reproducibility With the aim to promote open research, we release our collected datasets to the public, including all human annotations and preference rankings, enabling other researchers to build upon our work. We further provide comprehensive methodological documentation to ensure full reproducibility and transparency.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Giovanni Bonisoli, Maria Pia di Buono, Laura Po, and Federica Rollo. 2023. Dice: A dataset of italian crime event news. In *SIGIR '23: The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, July 23 - 27, 2023*. ACM.

JM Chein, SA Martinez, and AR Barone. 2024. Human intelligence can safeguard against artificial intelligence: individual differences in the discernment of human from ai texts. *Scientific Reports*, 14(1):25989.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12763–12771. AAAI Press.

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. 2019. Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in Brief*, 25:104076.

Cristina Garbacea, Samuel Carton, Shiyan Yan, and Qiaozhu Mei. 2019. Judge the judges: A large-scale evaluation study of neural language models for online review generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3968–3981, Hong Kong, China. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Jimpei Hitsuwari, Yoshiyuki Ueda, Woojin Yun, and Michio Nomura. 2023. Does human-ai collaboration lead to more creative art? aesthetic evaluation of human-made and ai-generated haiku poetry. *Comput. Hum. Behav.*, 139:107502.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Miao Li, Eduard Hovy, and Jey Han Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. Association for Computational Linguistics.

Zaijing Li, Gongwei Chen, Rui Shao, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *CoRR*, abs/2401.06836.

David M Markowitz, Jeffrey T Hancock, and Jeremy N Bailenson. 2024. Linguistic markers of inherently false ai communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1):63–82.

Hannah Mieczkowski, Jeffrey T Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. Ai-mediated communication: Language use and interpersonal effects in a referential communication task. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–14.

MOP-LIWU Community and MNBVC Team. 2023. Mnbvc: Massive never-ending bt vast chinese corpus. https://github.com/esbatmop/MNBVC.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the ruatd shared task 2022 on artificial text detection in russian. *arXiv preprint arXiv:2206.01583*.

Tatiana Shamardina, Marat Saidov, Alena Fenogenova, Aleksandr Tumanov, Alina Zemlyakova, Anna Lebedeva, Ekaterina Gryaznova, Tatiana Shavrina, Vladislav Mikhailov, and Ekaterina Artemova. 2025. Coat: Corpus of artificial texts. *Natural Language Processing*, 31(1):150–175.

Wei Song, Kai Zhang, Ruiji Fu, Lizhen Liu, Ting Liu, and Miaomiao Cheng. 2020. Multi-stage pre-training for automated chinese essay scoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6723–6733, Online. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024a. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. *CoRR*, abs/2402.11175.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Ashraf Elozeiri, Saad El Dine Ahmed El Etter, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 244–261, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Rustem Yeshpanov, Pavel Efimov, Leonid Boytsov, Ardak Shalkarbayuli, and Pavel Braslavski. 2024. KazQAD: Kazakh open-domain question answering dataset. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9645–9656.

Appendix

A Related Work

We first review studies investigating human evaluation of detecting MGT, and then summarize distinguishable features between human-written and machine-generated text, and end up with a discussion of relationship between individual factors, human detection capability and preference.

A.1 Human Detection on MGT

Early studies probing the outputs of less advanced generative models found that human evaluators could detect the differences. For example, Garbacea et al. (2019) reported that individual evaluators obtained 66.61% accuracy when evaluating online product reviews generated by 12 sequential models before 2019, and majority voting of five annotators' predictions improved to 72.63%. Ippolito et al. (2020) found that trained evaluators were able to detect GPT2-large text 71.4%.

However, since the emergent of GPT-3, it has become challenging for human evaluators to distinguish between human-written and machine-generated text. Evaluators could guess GPT-3-generated stories, news articles and recipes with 50% accuracy (Clark et al., 2021). 203 crowdsoured workers employed from Prolific show 57% average accuracy when discerning news generated by GPT-3.5-turbo and news collected from trusted news organizations, e.g., New York Times, Wall Street Journal, and they obtained 78% on social media comments, both with a wide spread in performance across individuals (Chein et al., 2024). Guo et al. (2023) also showed large differences in detection capability between experts and laymen, where the former is much higher than the later across different domains. Average accuracy of experts is above 90%, but random guess for amateurs, i.e., 48% and 54% for English and Chinese respectively. With expert annotators, academic paper abstracts are harder to detect than the Reddit answers. Detection accuracy ranges from 41 to 94 for Reddit with an average of 77, while 72 for abstracts with individual performance ranging from 60 to 84, revealing that scientific materials are more challenging even for experts who are researchers and frequent users of LLMs (Wang et al., 2024b).

In addition to binary detection human vs. AI, Wang et al. (2024a) performed 4-class detection to identify which LLM generated a given text, achieving accuracy of 21.2%, below the random guess of 25%, though annotators are all NLP PhD students who are exploring machine-generated text in their research. Dugan et al. (2023) asked annotators to select boundary sentences between human- and machine-authored snippets given a mixed collaborative text by a game tool. They found that players correctly identified the boundary sentence 23.4% of the time (chance being 10%).

As shown in Table 1, previous studies primarily focus on English and outputs of GPT-3.5-turbo, non-English languages and more advanced models have been under-investigated. The accuracy of human detection is influenced by languages, generative models, and domains of the given text, as well as the annotators. Large variance is observed among individual evaluators, particularly between experts and laymen. Therefore, we aim to fill this gap by a comprehensive case study exploring whether human can safeguard against AI over 16 dataset spanning nine languages, nine domains and 11 SOTA LLMs in four evaluation setups.

A.2 Distinguishable Signals

To characterize the specific linguistic properties that distinguish human from AI texts, Markowitz et al. (2024) compared real human hotel reviews to a set of LLM-created fake hotel reviews, and observed that machine texts had a more "analytical" style and exhibited increased use of affective language (stronger positive "emotional tone"). Similar work has shown that AI "smart replies" likewise demonstrate an emotional positivity bias (Mieczkowski et al., 2021). Guo et al. (2023) summarized that GPT-3.5-turbo writes in an organized manner, with clear logic, tending to offer a long and detailed answer. Also, the model responses generally strictly focused on the given question, whereas humans' are divergent and easily shift to other topics. Model answers are objective, typically formal, while humans' are more subjective, colloquial. and emotion-rich. Wang et al. (2024a) reported that human-written texts have some imperfect formatting patterns compared to machine text, e.g., initial double newlines, initial space, completely missing new lines in the paragraph, typos, inconsistencies within texts, or specific references and URLs.

In addition to probe the upper bound of human detection ability, we also try to answer: do these linguistic patterns mentioned above generalize to other languages and advanced AI systems? Can

they guide more accurate judgment of the text origins?

A.3 Detection Accuracy and Preference

What factors may influence individuals' discerning accuracy? Chein et al. (2024) investigated whether psychological attributes (e.g., fluid intelligence, executive functioning, empathy) or experience (e.g., smartphone and social media habits) predict the ability to differentiate between human and AI materials. Results show that only fluid intelligence — the capacity to reason and solve novel problems without relying on previously acquired knowledge, was strongly associated with the ability to differentiate human from AI materials. Other three factors were not significantly related to the overall detection accuracy.[1] They also find that increased exposure to mixed AI and human materials online does not enhance one's ability to differentiate human from AI materials, but rather, may make AI-generated content appear more human-like.

Combine together, language proficiency, cognitive intelligence, and familiarity with AI outputs may affect detection accuracy.

Intuitively, a high level of language proficiency will make it easier to spot overly formal, repetitive, or mechanically structured phrasing typical of AI outputs. Native speakers with strong awareness of cultural nuance can better understand cultural and idiomatic expressions, facilitating identify content that lacks human-like subtlety or creativity. Strong analytical skills can help detecting inconsistencies, logical flaws, style patterns that indicate MGT. Individuals with frequent interaction with LLMs may recognize their writing patterns like positive tone bias (Guo et al., 2023; Chein et al., 2024). Therefore, to probe the upper bound of human capability of detecting current SOTA LLM generations, all annotators for MGT detection and preference labeling are native speakers for each language, Masters', who are PhD and postdoc students studying LLMs, with 5 female and 14 male.

B Annotation Tool

To mitigate potential labeling biases arising from raw spreadsheet annotation and to enhance efficiency, we implemented two methods with optimized interfaces and workflows for our annotation: (1) a custom pipeline using the Google Workspace suite, including Apps Script, Google Sheets, and Google Forms. The core idea was to store all data in Google Sheets, use Apps Script to extract data and generate a survey in Google Forms;[2] and (2) Label Studio, an open-source multi-type data labeling and annotation tool with a standardized output format. We designed a custom template for our annotation task and collected results using this platform. The annotators were given the choice to use their preferred tool.[3]

C Datasets

C.1 Arabic

We collected three categories of text for Arabic including tweets, summaries, and news, involving four datasets.

Arabic Dialect Tweets We randomly selected 300 tweets from the QCRI Arabic POS Dialect dataset[4], which includes tweets in four dialects: Egyptian, Moroccan, Gulf, and Levantine. The dataset was originally curated for part of speech segmentation and covers a broad range of topics. The selected tweets were cleaned by removing the usernames of the original tweets' authors and parsing the words from each tweet to organize into sentences.

We generated 600 machine-generated tweets: 150 per dialect using GPT-4o and Qwen-2 (7.5B). Instead of calling APIs, we collected generations by the chat interface in order to esaily observe the patterns of the machine-produced content. The prompt employed for tweet generation was: *write a random tweet with dialect.* In this structure, `dialect` denotes one of the four targeted dialects. Additionally, the corresponding Arabic prompt was also used in generation:

اكتب تغريدة بال {dialect}

EASC Articles Summaries We used the Essex Arabic Summaries Corpus, a dataset featuring 153 Arabic articles sourced from Wikipedia, Alrai Newspaper, and Alwatan Newspaper. Each article is accompanied by 5 human-written summaries from which we sampled one. The corpus

[1] Fluid intelligence involves the ability to think abstractly, analyze patterns, make decisions in unfamiliar situations, and adapt to new challenges. Fluid intelligence is typically thought to be influenced by genetic and biological factors, and it tends to peak in early adulthood before gradually declining with age. It's often contrasted with crystallized intelligence, which is based on accumulated knowledge and experience.

[2] https://developers.google.com/apps-script

[3] https://github.com/HumanSignal/label-studio/

[4] https://huggingface.co/datasets/QCRI/arabic_pos_dialect

Language	Source/ Domain	Data License	Total Human	Sampled Parallel Data Human	GPT-4o	Claude	LLM2 Name (#)	Total
Arabic	Dialect Tweet	Apache 2.0	1400	300	300*	–	Qwen2 (300*)	900
	EASC	cc-by-sa-3.0	765	153	153	–	–	306
	Youm7 News	—	21,000	1,000	1,000	–	AceGPT (1,000)	3,000
	SANAD	cc-by-4.0	194,797	100	100	–	–	200
Chinese	Zhihu-QA	cc-by-4.0	224,761	588	588	–	Qwen-turbo (588)	1,764
	Student essay	cc-by-4.0	93,002	600	–	300*	ChatGLM4 (300*)	1,200
	Student essay	cc-by-4.0	51	51	–	51	ChatGLM4 (51)	153
	Government Report	MIT	200,409	500	500	–	Baichuan2-13B (500)	1,500
English	Peersum (Li et al., 2023)	cc-by-sa-4.0	5,158	400	200	200	–	800
Hindi	News	cc-by-4.0	3,995	600	600	–	–	1,200
Italian	DICE	cc-by-sa	10,518	300	300	Anita (300)	Llama3-405B (300)	1,200
Japanese	News	cc-by-nc-sa-4.0	7,110	300	300	–	–	600
Kazakh	Wikipedia	cc-by-sa-4.0	4,827	300	300	–	–	600
Russian	News	MIT	800,000	300	300	–	Vikhr-Nemo-12B (300)	900
	Academic summary	MIT	31,000	300	300	–	Vikhr-Nemo-12B (300)	900
Vietnamese	Wikipedia	cc-by-sa-3.0	600	600	600	–	–	1,200
	News	Kaggle data	290,282	600	600	–	–	1,200
Total	–	–	1,690,803	6,992	6,141	851	3,639	**17,623**

Table 4: Statistics of multilingual data for human annotation. Machine data with * means non-parallel data.

has a diverse range of topics, including art, education, politics, health, science, and finance. To generate machine-produced counterparts, we utilized GPT-4o using the prompt below. The prompt emphasizes that the generated summary should be highly informative, reserving all key ideas while remaining concise and to the point.

قم بتلخيص هذا المقال محافظا على اهم النقاط
ملتزما الايجاز و الدقة: {article} .

Youm7 News Articles The Kalimat dataset consists of articles sourced from the Egyptian news journal *Youm7*. This dataset comprises a total of 21,000 articles across various topics, providing a diverse base for text generation and analysis. For our study, we sampled 1,000 news and generated using GPT-4o (a multilingual LLM) and AceGPT (a Arabic-centric LLM).

The text generation process was guided by a carefully crafted prompt to simulate the typical writing style of *Youm7* articles. This prompt instructed the models to act as a professional journalist, ensuring coherence and alignment with the original content's structure and tone. Below is the prompt used for this task:

تصرف كأنك كاتب أخبار محترف، واكتب مقالًا
باللغة العربية يتألف من حوالي
{word_count_rounded} كلمة، بعنوان {title} .
اعتمد في أسلوبك على طريقة كتابة مؤلفي
{source} ، مع الأخذ في الاعتبار أن الموضوع
يندرج تحت فئة {topic_arabic} .

The goal of the generated samples was to maintain narrative consistency and topical relevance while adhering to the typical article patterns found in *Youm7*.

SANAD News The SANAD news dataset (Einea et al., 2019) is an extensive collection of Arabic news articles in Modern Standard Arabic (MSA), designed to support various Arabic NLP tasks. SANAD dataset comprises over $190,000$ articles from three prominent news websites: AlKhaleej, AlArabiya, and Akhbarona, and it is released under the license of CC-BY-4.0.

For this study, we generated a total of 500 news articles using titles from the SANAD news dataset, with 100 articles for each of five selected news category topics present in SANAD spanning finance, politics, sports, medicine, and technology. The articles were generated using the OpenAI's GPT-4o and the prompt used to generate these articles is shown below:

تصرف كأنك كاتب أخبار محترف ، و اكتب مقالًا
باللغة العربية يتألف من حوالي
{word_count_rounded} كلمة، بعنوان {title} .
اعتمد في أسلوبك على طريقة كتابة مؤلفي
{source} ، مع الأخذ في الاعتبار أن الموضوع
يندرج بشكل عام تحت فئة {topic_arabic} .

In this prompt, we ask the LLM to act as a professional news writer, and write an Arabic news article consisting of around '`word_count_rounded`' words carrying the title '`title`', using the writing

style of authors of 'source' while taking into account that the topic falls under the general umbrella of 'topic_arabic'.

'word_count_rounded' represents the number of words rounded to the nearest hundred of the corresponding human-written article for the same title. 'source' is the publisher name of the given title obtained from SANAD. 'topic_arabic' is one of the five topics mentioned earlier selected to match the topic of the given title.

C.2 Chinese

We collected Zhihu question answering, high school student essays and government reports.

Zhihu QA Based on a Chinese question answering dataset collected from Zhihu, where the users can post and answer questions as well as liking or sharing them, we randomly sampled 588 examples.[5] They span life, technology, art, science, emotion, law, fashion, humor, and China-specific cultures. For emotion-rich questions, we specially collected across 10 topics including breaking-up, quarrel, happiness, marriage issue, divorce, depression, inferiority, parents, forgiveness, future concern (10 questions for each), emphasizing emotional and humorous responses. Then we collect corresponding responses from a multilingual LLM GPT-4o and a Chinese SOTA model Qwen-turbo.

Junior and Senior High School Student Essays We sampled 600 Chinese student essays from Song et al. (2020), attempting to ensure the uniform distribution over the genre of an essay and the grade of the student.[6] However, the dataset does not provide the corresponding problem statements or titles of essays, making it challenging to generate the counterpart by LLMs. To produce the high school student essays by LLMs, we approached by collecting: (i) essay problem statements of Gaokao from 1977 to 2024, along with some junior high school problems, in total of 376 problems, and (ii) 51 (problem, essay) pairs, with the excellent essays published in the website as the human text.[7] We sampled 300 problem statements from (i), prioritizing the latest years, and the whole set of (2), and then generate machine counterparts by Claude-3.5-Sonnet and ChatGLM-4-9B-chat.

Government Report The GovReport dataset, sourced from the MNBVC GovReport subset (MOP-LIWU Community and MNBVC Team, 2023), includes titles and main bodies of reports from various entities, such as schools, corporations, and local government units, without additional labels or classifications. We randomly sampled 500 entries to generate a diverse, high-quality LLM dataset. To achieve varied outputs, we used a set of eight distinct prompts, encompassing continuation tasks based on titles or opening sentences and rephrasing tasks based on the full report content. These prompts, presented in Chinese, were alternated to guide LLMs in generating work reports, with responses collected from GPT-4o, Baichuan2-13B-Chat, and ChatGLM3-6B.

C.3 English Peer Meta Review

We randomly sample 600 reviews (NeurIPS 2021-2022) from PeerSum dataset (Li et al., 2023). Peersum crawled both meta reviews and reviews from each reviewer and rebuttal from authors from openreview[8]. We leveraged this dataset to generate meta reviews based on comments from other reviewers and compared with human-written meta reviews. We generate meta reviews using `GPT-4o-2024-08-06` and `claude-3.5-sonnet-20240620` respectively, with prompt "Generate a meta review based on the reviews' opinions and authors' rebuttal to make the final decision on whether the paper should be accepted: `{Reviews}`\n\n Meta review:", where "`Reviews`" contains both the reviewers' review and authors rebuttals. We sampled 400 machine-generated meta reviews, paired with humans, and asked annotators to discern.

C.4 Hindi News

The BBC Hindi news article dataset comprises a diverse selection of around 4,000 Hindi news articles sourced from the BBC Hindi website, covering a wide range of topics and categories. Each article includes three primary components: headline, main content and the thematic category of the article.

We sampled 600 articles from the original dataset, focusing on articles that span multiple

[5] https://huggingface.co/datasets/zirui3/zhihu_qa

[6] Each instance includes the essay, and the genre (character, narratives, scenery, objects, argumentative and prose) and the rating (bad, moderate, good, excellent) of the essay, and the grade of the student (7-9 is junior and 10-12 is senior high school).

[7] https://docs.google.com/spreadsheets/d/1iwtnQkamxoqnThWUT9F007WUHm1XmNESAwSVFM8xNdU/edit?usp=sharing

[8] https://openreview.net/

prominent topics in Indian news to ensure a balanced representation. These articles were selected randomly to avoid any thematic bias and provide a broad scope for comparison. We generated 600 machine-written samples using GPT-4o model, employing a prompt that guided the model to write concise and formal articles: "Here is a news headline: 'headline' and the content: 'content'. Write a machine-generated version of the news based on this headline".

C.5 Italian DICE News

For Italian, we used DICE (Bonisoli et al., 2023) — local news from La Gazzetta di Modena ($\approx 10,000$ samples) licensed as CC-by-nc-sa. We sampled 300 news, where the original text has at least 1000 characters. We applied three large language models: Llama-3.1-405b-instruct, GPT-4o, and Anita (an Italian fine-tuned Llama-3.1-8B: Polignano et al. (2024)).

When generating in Italian, we use one of 4 possible system prompts as Table 5. The model is always prompted as a journalist but the mother tongue and the newspaper scope may vary lightly. The general format is the following:

> **System:** You are an Italian journalist writing for a national newspaper focusing on criminal events happening in the area surrounding Modena.
> **User:** Write a piece of news in Italian, that will appear in a local Italian newspaper and that has the following title: ...

C.6 Japanese News

We randomly sample 300 news articles of the BBC news from the XLSUM dataset (Hasan et al., 2021). XLSUM has title, content, and summary of a news article. We generate the corresponding news content using `GPT-4o-2024-08-06` based on the article titles. For both text generated with GPT-4o and the human-written text, we remove obvious formatting indicators (e.g., line breaks and template messages at the beginning and end of the texts).

We generate 75 articles for each of the four settings: simple prompt zero-shot, diverse expression zero-shot, content-rich zero-shot, and few-shot. Simple prompt zero-shot, diverse expression zero-shot, and content-rich zero-shot use the following instructions, respectively: "次のニュースタイトルに合わせたニュース記事を生成してください。" (Please generate a news article that matches the following news title.), "次のニュースタイトルに合わせたニュース記事を多様な表現を使用して生成してください。" (Please generate a news article that matches the following news title, using diverse expressions.), and "次のニュースタイトルに合わせたニュース記事を生成してください。このとき生成するニュース記事には、誰かに対するインタビューや実際の出来事を組み込んでください。" (Please generate a news article that matches the following news title. When creating the article, include interviews with individuals or actual events.). The few-shot uses the same instruction as the simple prompt zero-shot. Add "ニュースタイトル:" (news title:) and "ニュース記事:" (news article:) at the beginning of the news title and the content, respectively. We randomly sample three contents as examples for few-shot, and use the same examples for all generations.

C.7 Kazakh Wikipedia

KazQad is a closed question-answering dataset focused on the Kazakh language (Yeshpanov et al., 2024). It contains 5,000 distinct passages covering 1,700 topics derived from Kazakh Wikipedia. These passages span a variety of domains, including art, science, history, sports, and other general topics.

We randomly selected 300 titles along with their corresponding paragraphs, ensuring that each paragraph contained at least 3-5 sentences. Additionally, the sampled texts were cleaned and merged by title to increase the length of the texts. This step was necessary because some samples in the original dataset contained extraneous elements such as references, markdown formatting symbols, and other unnecessary characters. For the machine-generated data, we used the GPT-4o model to generate passages based on the sampled titles. The generation process was initiated with the following prompt: "Please, write one paragraph about the following topic in Kazakh: [title]."

C.8 Russian

We generated machine text for both news and summaries for Russian.

News Based on Lenta.ru news from Corus with around 800,000 news samples, we sampled 50 cases for each topic among the six topics: Russia, World, Economy, Sport, Culture, and Science & Technology. We prompted GPT-4o and Vikhr-

English-first LLMs	
System Prompt	
You are an Italian journalist writing for a national newspaper focusing on criminal events happening in the area surrounding Modena	You are an Italian journalist writing for a local newspaper focusing on criminal events happening in the area surrounding Modena
You are an Italian-French journalist writing in Italian about criminal events happening in the area surrounding Modena	You are an Italian-American journalist writing for a local newspaper focusing on criminal events happening in the area surrounding Modena
User Prompt	
Write a piece of news in Italian, that will appear in a local Italian newspaper and that has the following title:	Write a piece of news in Italian, that will appear in a local Italian newspaper and that has the following title:
Write a piece of news in Italian, that will appear in a national Italian newspaper and that has the following title:	Write a piece of news in Italian, that will appear in a local Italian newspaper and that has the following title:
Italian-first LLMs	
System Prompts	
Sei un giornalista italiano che che scrive per un giornale nazionale focalizzandosi su eventi criminali che accadono a Modena	Sei un giornalista italo-francese che scrive in italiano su eventi criminali che accadono a Modena
Sei un giornalista italiano che scrive per un giornale locale focalizzandosi su eventi criminali che accadono a Modena	Sei un giornalista italo-americano che scrive per un giornale locale focalizzandosi su eventi criminali che accadono a Modena
User Prompts	
Scrivi un articolo di giornale in italiano. L'articolo sarà pubblicato su un giornale locale e avrà il seguente titolo:	Scrivi un articolo di giornale in italiano. L'articolo sarà pubblicato su un giornale nazionale e avrà il seguente titolo:
Scrivi un articolo di giornale in italiano. L'articolo sarà pubblicato su un giornale locale e avrà il seguente titolo:	Scrivi un articolo di giornale in italiano. L'articolo sarà pubblicato su un giornale locale e avrà il seguente titolo:

Table 5: Italian machine generation prompts.

Nemo-12B-Instruct-R-21-09-24 to generate to corresponding machine text using "Напиши новость в области "topic"с сайта lenta.ru используя заголовок title. Ты должен генерировать новость без заголовка. Новость: "

Academic Article Summaries Based on ai-forever/ru-scibench-grnti-clustering-p2p with 31 000 samples, we sampled 30 summaries for each topic among 'Psychology', 'Mechanical Engineering', 'Agriculture and Forestry', 'Geology', 'Biology', and 'Energy', and generated machine counterparts using GPT-4o and Vikhr-Nemo-12B-Instruct-R-21-09-24. We used the following prompt: "Напиши краткое содержание для статьи в области "topic"используя заголовок title. Ты должен генерировать краткое содержание без заголовка. Содержание:"

C.9 Vietnamese

Vietnamese News The dataset is crawled from Lao Động newspaper[9] before May 2022 (before the releases of all LLM that support Vietnamese). The dataset contains 290,282 articles with a headlines and a summary of an article in various topics, such as politics, lifestyles, legal, etc. We cleaned the data by removing some missing values and too short summaries (less than 20 words), then randomly selected 600 examples to generate the corresponding 600 machine-generated summaries with GPT-4o using title of articles. The prompt for generation process was

> *Bạn là một nhà báo Việt Nam chuyên viết những mẩu tin tóm gọn cho các bài báo*

[9] https://www.kaggle.com/datasets/phamtheds/news-dataset-vietnameses

Dialect	Human	GPT-4o	Qwen2-7.5B	Overall MGT
EGY	52.00	53.33	58.67	56.00
MOR	54.00	53.33	48.00	50.67
LEV	69.33	14.67	58.67	36.00
GULF	81.33	26.67	30.67	28.67

Table 6: Arabic dialect tweet human detection accuracy over human vs. GPT-4o vs. Qwen2-7.5B. Machine-generated text is harder than human text to discern. GPT-4o is harder than Qwen2.

> *bằng cách sử dụng tiêu đề của chúng. Hãy viết cho tôi một đoạn tóm tắt bài báo tiêu đề dưới đây:*

which means *You are a Vietnamese journalist who writes summaries for articles using their headlines. Please write me a summary of the article with the following headline:*

Vietnamese Wikipedia We randomly crawled 600 sites from Vietnamese Wikipedia[10] with its ID, title, and the introduction part of each topic. For our studies, we generated 600 introduction given a subject using GPT-4o with the prompt:

> *Bạn là một nhà đóng góp cho Wikipedia tiếng Việt. Hãy viết cho tôi một đoạn giới thiệu ngắn gọn bằng tiếng Việt về chủ thể bên dưới để đăng trên trang Wikipedia. Lưu ý bắt buộc viết với khoảng* word_count *từ.*

which means *You are a contributor to Vietnamese Wikipedia. Please write me a brief introduction in Vietnamese about the subject below to post on the Wikipedia page. Note that it must be written within* word_count *words:*

In this prompt, to avoid the newly generated text has too long passage, since GPT-4o tends to write much longer than human does for a Wikipedia topic, we set the word limits word_count as word count of the original one.

D Distinguishable Signals

This section elaborates detection setting, accuracy and distinction signals summarized by 18 annotators for 16 datasets one by one.

D.1 Arabic

Arabic Dialect Tweets The annotation was conducted under the setting III. Single-binary, where a total of 900 tweets across four dialects were analyzed, with ratio of machine-generated vs. human-written text as 2:1 (GPT-4o and Qwen2-7.5B).

The annotators were not very confident about their choices when deciding whether the data is human-written or not. They achieved an overall accuracy of 50.06%, highlighting the difficulty of distinguishing machine-generated content in short texts. Overall, 64.17% of human-written tweets were correctly annotated, 36% human-written tweets were mis-identified as machine-generated. Table 6 presents the human detection accuracy on human text and machine-generated tweets across four dialects, machine-generated text is harder than human text to discern. GPT-4o outputs are more similar to human text, thus more difficult to distinguish compared to Qwen2.

Since LLMs effectively replicate native speaker vocabulary, human detection cannot rely on distinguishing lexical distinctions. A key indicator of machine-generated text was the use of emojis in an overly formulaic manner. Additionally, some tweets addressed topics that would typically be considered trivial or unlikely for humans to post, further suggesting machine generation. Machine-generated tweets also exhibited unnatural tones, misused native expressions, or contained incomplete content. A notable issue in Qwen2's output was the inclusion of words from other languages within sentences and the mixing of dialects, which is uncommon in natural usage.

EASC Articles Summaries We sampled 100 (human, GPT-4o) pairs to identify which text is written by human, achieving 82% accuracy. The annotator differentiated human-written text from the machine-generated text based on five empirical distinguishable signals.

- **Informative:** LLM summaries are more informative, while humans may overlook key points due to emotional biases, personal perspectives, or incomplete understanding of articles.
- **Abstract and concise:** LLMs present ideas in a more abstract and concise way, without emphasis on specific points. Human summaries often inject personal opinions and beliefs.
- **Religious language:** Humans can accurately use religious language, whereas LLMs generate text in standard language.
- **Prompt reflection:** LLMs often start with

[10] https://vi.wikipedia.org/

words from the prompt, such as

هذا المقال

- **Formatting hints** Human-written summaries contain typos or grammar errors. Machine-generated text includes markdown elements, making it more easily detectable as machine-generated.

Youm7 News Articles We paired 1,000 (human, GPT-4o) examples to identify which text is written by human, achieving accuracy of 92.7%. A key finding was that human-written paragraphs were typically presented as a single continuous block of text, whereas GPT-generated article was segmented into smaller sections. This structural difference served as a crucial cue for identifying machine-generated content. No significant variations were observed in thematic patterns or narrative sequences, paragraph segmentation emerged as a key indicator, highlighting the role of structural features in detection.

SANAD News Using the SANAD Arabic news dataset in Modern Standard Arabic (MSA), we evaluated 100 samples to identify human vs. GPT-4o text, achieving 100% annotation accuracy under the setting I. Pair-binary. We identified several key features that differentiate machine-generated text from human-written text.

- **Markdown presence:** Machine-generated text often contains markdown, which is absent in human-written text.
- **Formatting style:** Machine-generated text is consistently formatted into structured paragraphs, whereas human text tends to appear as large, unstructured blocks of text.
- **Content density:** Human text is richer in factual information, while machine-generated text includes generic statements. For instance, phrases such as mentions of "Saudi Vision 2030" are more frequent in machine-generated text.
- **Source attribution issues:** Machine-generated text often contains source attributions, but these are sometimes incorrect or against the prompt.
- **Presence of numbers and specific meta data:** Human-written text contains more supporting numerical details, including URLs, phone numbers, currency exchange rates, dates, and amounts, which are not as frequent in machine-generated text.
- **Use of English terms:** Human text includes sporadic English terms, especially in specialized contexts, while machine text remains fully in Arabic.
- **Hashtags and social media elements:** Hashtags are commonly found in human-written text, whereas machine-generated text lacks such social media elements.
- **Narrative structure:** Human-written text follows a narrative structure with a sequence of events, dates, and timelines. On the other hand, machine-generated text tends to resemble an essay on a given topic rather than a chronological news report.
- **Formatting consistency:** Human text formatting varies significantly, often appearing as inconsistent or messy blocks of text with irregular spacing or newlines. We attribute this apparent inconsistency to the fact that the human texts are composed by a large number of authors, while machine text is written by a single model which provides text that is more polished and consistent.
- **Readability and grammar:** Machine-generated text is generally more readable and grammatically correct, while human text may contain stylistic inconsistencies.

D.2 Chinese

Zhihu QA The detection involves six unique individuals, with three female and three male annotators. We pair (human, GPT-4o) and (human, Qwen-turbo), and annotate both under setting I, achieving the average accuracy of 1.0 and 0.98 respectively. The distinguishable factors are summarized below.

- **Humans share personal experience and feelings.** For emotion-rich questions, humans provide empathic comforts by sharing their real personal experiences. However, narratives or stories generated by models tend to appear contrived and artificial, making it easy for humans to detect that they are not grounded by facts, similar to stories created by kids. Model responses thus typically offer solutions, but they are theoretically sound, while often lack practical applicability.

- **Human answers can be informal, mean and sharp.** Human responses sometimes are mean, strongly opinionated, and influenced by personal biases, reflecting a more self-centered perspective. LLMs provide general and less engaging information, but often attempting to help users and offering problem-solving assistance. Some individuals appreciate human answers for their authenticity and directness, others may find them offensive.
- **Different Intent.** Zhihu users prioritize expressing their feelings and opinions, rather than assisting and directly addressing the seeker's needs. So they often do not respond to the question directly. Though LLMs aim to assist but typically offer general information related to the entities or concepts in the question, rarely responding with direct and sharp answers.
- **Human answers can be extremely short or long.** Some human answers are excessively long and lack proper paragraph segmentation, making them harder to read. LLM-generated responses are generally well-segmented and structured with bullet points or listed points with bold subtitles.
- **LLM responses lack deviation.** LLMs adhere rigidly to instructions, presenting limited flexibility in their responses. When prompted to answer emotion-rich questions with greater empathy, their generative patterns remain predictable, often relying on superficial expressions such as more emoji inserted instead of deeply integrating empathy into the content. For instance, responses like *I'm sorry this happened to you; you should probably consult a professional*, offer minimal support and feel unhelpful.
- **Other indicators** Human answers sometimes contain timeline of answer update and references.

Improved Prompts We carefully designed prompts using a few different user personas: one with a positive attitude, another with a rational and realistic outlook, and a humorous one. For emotion-rich questions, we also applied GPT-4o utilizing ECoT, a framework designed to enhance LLMs' emotional and empathetic responses (Li et al., 2024). However, the responses are often brief, with emojis but minimal information, which limits their usefulness.

High School Student Essays We perform human detection for student essays in two settings. One NLP postdoc who is a Chinese native speaker did the detection under the setting I for 102 parallel pairs (hwt, mgt_1) and (hwt, mgt_2), obtaining the accuracy of 98%. For 600 non-parallel cases under the setting II. Pair-four-class, with 150 pairs for each class, another NLP postdoc and two NLP PhD students participated in annotations. They achieved an accuracy of 0.96, 0.96 and 0.99 respectively.

During the annotation of high school student essays, we identified four remarkable distinction signals that make machine-generated texts easily recognizable:

- **Title:** From the perspective of formatting, machine text tends to begin by "title: 《xxx》 or title: xxx", and have newlines between paragraphs while humans does not have.
- **Formulaic structure:** From the structure and the content of essays, the structure of machine-generated essays is generally formulaic. They often adopt an argumentative style that begins paragraphs with phrases such as "first, then, moreover, additionally, finally, and overall" (i.e., 首先，其次，然而，总之，最后), while HWT is more flexible and the styles are more diverse. Some MGT even use bullet points in an essay, which is rare in human text.
- **Sentiment:** Machine-generated essays tend to adopt a more neutral or positive tone, generally expressing less emotion compared to human-written essays. While real students often express a range of emotions, including sadness, anger, confusion, and a sense of being lost. These emotions reflect the authentic feelings of young people at that age.
- **Style:** Machine-generated content may incorporate elements from other genres, such as official document styles. Additionally, machine-generated content may unexpectedly switch to multilingual content, for example, outputting an English paragraph during Chinese content generation.

Original vs. Improved Prompt Based on the findings above, we further refined the prompts in the following ways: (1) instructing the model to avoid outputting titles, as these often serve as clear detection cues; (2) discouraging excessive use of connecting words like 'first of all,' 'secondly,' 'then,'

and 'finally'; and (3) preventing the mixing of languages other than Chinese.

Government Report Under setting IV. Triplet-three-class, we presented human-written texts vs. two model outputs across 500 samples to a native Chinese-speaker (postdoctoral researcher specializing in NLG), and asked the annotator to identify which text is human-written. The annotator achieved an accuracy of 97.2%. Human-authored texts were typically longer and contained richer details, often covering multiple topics, while machine-generated texts were generally shorter, lacking noticeable rhythm variations. Additionally, machine-generated texts occasionally included distinct symbols, such as bold formatting and English words. Errors in distinguishing humans from machine-generated texts primarily occurred when the machine-generated outputs were of similar length to human-written ones.

D.3 English Peer Meta Review

Human detection accuracy is 99.75%. Before seeing labeled samples, the annotator found the distinction was obvious. After reviewing a few examples, the annotator was extremely confident in distinguishing human-written from machine-generated text based on indicative features of MGT.

Machine-generated peer reviews exhibited distinct, predictable patterns. They frequently followed a structured format, often providing explicit decisions as "Decision: Accepted", which was rare in human-written reviews. MGT also commonly used headings such as "Strengths" and "Weaknesses". MGT tended to explicitly summarize multiple reviews while the human reviewers rarely did. An example of MGT is like "Based on the reviews, this paper presents methods for embedding numerical features to improve deep learning models for tabular data. The key points are:".

MGT additionally presents a notable preference for bullet points, even when a paragraph format was feasible, with minimal variation in bullet style. MGT responses tend to be more uniform in length and generally longer than human reviews. In contrast, human-written meta reviews lacked a fixed structure, rarely used bullet points, and exhibited greater variation in length, though they were typically shorter than MGT.

D.4 Hindi News

An overall accuracy of 85.17% was achieved in distinguishing between the machine-generated and human-written Hindi BBC news articles. The analysis reveals key stylistic and content-based distinctions. One significant observation was that machine-generated news content tended to be more concise, presenting less overall information compared to its human-written counterpart. Additionally, machine-generated articles included fewer mentions of names of persons and specific dates, elements that are typically embedded within human-authored news to enhance credibility and specificity.

Furthermore, stylistic disparities were evident, particularly in the absence of colloquial language elements commonly used by human authors. Human-written articles often incorporate idioms, culturally significant phrases, and even a blend of Urdu vocabulary, reflecting the linguistic diversity and nuance of the Hindi language. These elements, however, were noticeably missing in machine-generated content, which instead adhered strictly to the main topic, with minimal linguistic or thematic deviations. This adherence to topic, while adding to clarity, lacked the depth and regional authenticity often present in human-created Hindi news content. Also there was use of English text in the machine generated version.

D.5 Italian News

Based on human text from DICE dataset (Bonisoli et al., 2023), we randomly selected machine-generated text from GPT-4o, Anita (Polignano et al., 2024) and Llama3-405B (Dubey et al., 2024) outputs, resulting in total of 300 (human, MGT) pairs. A native Italian speaker was asked to choose which text is human-written. Detection accuracy is 88.0% for Anita, 99% on Llama3-405B and 100% for GPT-4o.

We identified the following distinguishable signals. When generating news, all models especially the larger ones, Llama-405b and GPT-4o have consistent formatting, e.g., article title is always enclosed in markdown-like double asterisks "**", the city where the issue happens is always mentioned first. These cues are immediately recognizable. After seeing only 2-3 examples, human annotator was confident to make decisions and obtain close to 100% detection accuracy.

D.6 Japanese News

We divided 300 pairs into two groups and two annotators independently annotated them, to identify which text was a human-written BBC news article. Human annotators achieved an average accuracy of 62%. We identified several distinctive characteristics of LLM-generated texts:

- **Style:** Some texts failed to match news writing styles. For instance, in Japanese, writing typically either uses the *desu/masu* or *dearu* style. While news articles conventionally use the *dearu* style, LLM-generated texts often use the *desu/masu* style.
- **High reliance on headlines:** The opening sentences frequently relied on the title (headline), either through direct repetition or close paraphrasing.
- **Formulaic phrase:** LLM outputs sometimes contained formulaic phrases like *This article is provided as fiction* or *Here are five reasons: 1. ....*
- **Typos and grammar issues:** We observed various typographical errors and fluency issues, such as *This four has shocked the entire UK* or *An event that shook the air in Myanmar*.

These findings suggest that LLMs have not yet fully mastered the stylistic conventions of news writing. Explicit instructions that indicate the domain-specific format and style may help the LLM output.

D.7 Kazakh Wikipedia

Two native Kazakh speakers annotated 300 (human, GPT-4o) pairs under the setting I, with detection accuracy of 79.67%. We summarized the following features of MGT.

- The generated text lacks diversity in expression, often using repetitive sentence patterns and predictable phrasing, such as frequently ending with "bolyp tabylady", which contributes to a mechanical and formulaic feel.
- The generated text rarely includes concrete facts, such as numbers or years, and when they do appear, their occurrence is minimal.
- The text also frequently includes flattering language, which can be another signal of its artificial nature.
- Human-written texts include more Kazakh-specific cultural references where applicable, making them more relatable and authentic, which serves as a clear signal of human-generated content over LLM-generated text.

D.8 Russian

News Articles A native Russian speaker (NLP PhD student) annotated 300 examples under the setting I, where machine-generated text was randomly selected from either GPT-4 or Vikhr outputs. Accuracy is 100%. We found that human texts often include specific details like exact dates, numbers, names of people, places, or things (especially those not mentioned in the title or from ordinary backgrounds), as well as ages and other specifics. Additionally, human-written texts tend to reference sources, which is a strong indicator.

Academic Article Summaries A native Russian speaker (NLP PhD student) was asked to identify whether the given text is human-written or machine-generated under the setting III. Single-binary. A text is randomly chosen from human-written, GPT-4, or Vikhr outputs (300 samples). The overall accuracy is 80% (Psychology: 80.0%, Mechanical Engineering: 74.0%, Agriculture and Forestry: 74.0%, Geology: 86.0%, Biology: 86.0%, and Energy: 80.0%). There are several indicators. Machine-generated texts, especially from Vikhr, often begin with a paraphrased version of the title, which is a strong indicator. Human-written texts typically contain details such as numbers, references, and names. In some Vikhr outputs, sentences may lack coherence, and even after preprocessing, certain artifacts remain visible in the text.

D.9 Vietnamese

We annotated 600 news summaries and 600 Wikipedia introduction passages in the setting of I. Pair-binary. We obtained accuracy of 80.33% on news summaries, but random guess 50.67% on Wikipedia text, showing minimal differences between human-written and machine-generated Wikipedia passages.

A key finding was that, for news articles, human tends to provide more details to support the statement (such as date, location, and other specific information), while LLMs tend to offer a general summary. Besides, GPT-4o usually uses short sentences in summaries, but Vietnamese journalists usually write long sentences.

For Wikipedia passages, it is much harder for humans to distinguish since GPT-4o was well-

trained on Wikipedia data. The differences between human-written and machine-generated Wikipedia text is minor, but there are still some distinctions. For example, when GPT-4o was asked to write about Hai Phong city, city demographics and geography are expected, while it generated a review-like sentence: *Modern infrastructure, along with open investment policies, have helped Hai Phong become an attractive destination for domestic and foreign investors, contributing to the city's sustainable development.*

E Fill the Gap by Prompting?

In this section, for each dataset, we first present how we designed the improved prompts, and then elaborate whether the new generations are improved and fill the previously-observed gaps between humans. Original and improved prompts for all datasets and languages are summarized in Table 8, and the detection accuracy on new content and fill-gap survey results are demonstrated in Table 7.

E.1 Arabic

Tweets To enhance the quality of machine-generated Arabic tweets, prompts were refined to better capture human emotions and generate tweets that can reflect authentic human experiences. One such improved prompt was: "Generate a random tweet in Arabic. Use dialect dialect, use human emotions and experiences. Output the tweet only" and its corresponding Arabic translation:

اكتب تغريدة عشوائية باللغة العربية. استخدم اللهجة {dialect} وعبّر عن مشاعر وتجارب إنسانية

These adjustments addressed critical gaps appeared in prior outputs. The newly-generated tweets touched on relatable human topics and expressed genuine emotions tied to daily experiences. We used GPT-4o for generation as it yielded promising results, while Qwen is underperformed. As mentioned in Appendix D.1, Qwen texts often include non-Arabic words within the tweets.

To evaluate the effectiveness of the improved prompts, 200 sampled tweets were assessed by annotators, determining whether the outputs addressed the identified gaps. The evaluation revealed that 53% of the tweets fully met the criteria, 29% partially did, and 18% did not. However, some limitations were observed: GPT-4o frequently added irrelevant hashtags at the end of tweets, making the outputs identifiable as machine-generated, and the tweets often carried an **overly optimistic tone**. Even when negative experiences were mentioned, the sentiment leaned toward positivity and new beginnings. Overall, while these improved prompts succeeded in filling some previous gaps, new issues emerged, which could be addressed with more precise instructions, such as explicitly avoiding hashtags or adopting a more somber tone when required.

EASC Summary Building on the identified discrepancies between human-written and machine-generated summaries discussed in Appendix D.1, the generation prompt was refined to address these gaps, resulting in the following prompts:

قم بتلخيص المقال التالي محاولا محاكاة انسان له اراء و معتقدات فكرية او دينية ملتزما الدقة و الايجاز فمتوسط مستوى البشر :{article} .

The data was re-generated using GPT-4o, following the enhanced prompt. This enhancement led to a decrease in annotation accuracy, which dropped to 63% from 82%. However, there is still typical machine writing style in the new content, like the use of common machine-generated phrases (e.g., يتناول هذا المقال...), clearly indicating that the text was produced by a machine. Additionally, the newly-generated summaries exhibited several characteristics.

- The summaries seem to reflect a particular ideology, often emphasizing religious themes or beliefs, which influenced the tone and content of the generated text.
- The generated sentences are not concise, limiting the clarity and precision of the summaries.
- The language used to describe events or natural phenomena tendd to be more emotional, with a tendency to exaggerate or reflect personal preferences.

SANAD Before prompt engineering, the accuracy of annotating which of two texts is machine or human written was 100%. After prompt engineering, the accuracy dropped to 66% indicating that prompt engineering works at least partially to bridge the gap between the writing styles of humans and machines.

We follow the same model choices and data subset outlined in C.1 but we enhance the prompt as shown below:

Language	Source/Model	#Example	#Annotator	Yes	Partially	No
Arabic	Dialect Tweet	200	1	106	58	36
	ESAC Summary	130	1	82.0% → 62.7%		
	Youm7 News	200	1	92.7% → 52.0%		
	SANAD News	200	1	100% → 66.0%		
Chinese	Zhihu-QA	100	3	99.7% → 94.3%		
	Zhihu-QA	200	1	32	70	98
				86	36	78
	Student essay	200	3	84	36	80
				45	23	132
	Government Report	200	1	59	38	103
English	Peersum	200	1	2	113	85
Hindi	News	200	1	85.2% → 66%		
Italian	DICE News (Anita)	300	1	88% → 81%		
	DICE News (Llama3-405B)	300	1	99.7% → 84%		
	DICE News (GPT-4o)	300	1	100% → 85%		
Japanese	News	200	2	86.4% →86%		
Kazakh	Wikipedia	200	2	105	89	6
Russian	News	200	1	100% → 86.5%		
	Academic summary	200	1	80% → 69%		
Vietnamese	Wikipedia	600	1	50.7% → 47.3%		
	News	600	1	80.3% → 63.2%		
Total	–	4,730	25	**87.6% → 72.5%**		

Table 7: Human detection accuracy differences on original vs. improved generations, and survey distribution evaluating whether the new generations fill the gap: Yes, Partially or No.

تصرف كأنك كاتب أخبار محترف، واكتب مقالًا باللغة العربية يتألف من حوالي {word_count_rounded} كلمة، بعنوان {title} . احرص على أن يكون المقال غنيًا بالمعلومات والتفاصيل الدقيقة، متضمنًا أرقامًا، تواريخ، ومعلومات كمية إذا كانت مناسبة. تأكد من إدراج روابط، أرقام هواتف، أو أسعار صرف العملات عند الضرورة. استخدم بعض المصطلحات الإنجليزية المتخصصة إن كان السياق يتطلب ذلك، وتجنب الاكتفاء بعبارات عامة مثل »ويأتي هذا التطور في ظل توجه عالمي...« دون دعم بالأحداث أو التفاصيل. راعِ سرد الأحداث بترتيب زمني واضح يعكس التسلسل المنطقي للوقائع. احرص على أن يكون النص مكتوبًا في فقرة واحدة فقط بغض النظر عن طوله، دون تقسيمه إلى فقرات متعددة. اجعل التنسيق غير مثالي (مثل وجود مسافات غير متساوية) لتعكس طبيعة الكتابة البشرية، ولا مانع من استخدام الهاشتاجات عند الحاجة. يرجى مراعاة أن الموضوع يندرج بشكل عام تحت فئة {topic_arabic} . يرجى عدم استخدام Markdown نهائيًا.

Similar to the original prompt, this prompt instructs the LLM to act as a professional news writer and generate an Arabic news article of approximately '`word_count_rounded`' words with the title '`title`', ensuring that the article's general topic aligns with '`topic_arabic`', one of the five predefined categories.

The enhancements made by this prompt can be summarized as follows:

- Produces content rich in information and accurate details.
- Includes specific numbers, dates, and quantitative data where applicable.
- Encourages the incorporation of specialized English terms when the context requires it.
- Advises against vague phrases that are unsupported by concrete events or details.
- Requests a clear chronological narration of events.
- Requires the article to be written as a single paragraph, regardless of its length.
- Allows imperfections in formatting (e.g., uneven spacing) to mimic human-like writing.
- Permits the use of hashtags when appropriate.

- Prohibits the use of `Markdown`.

The following observations were made from the results of the enhanced prompt:

- **Fake phone numbers** (e.g., 123456789) indicate machine-generated content.
- **Correct phone numbers or URLs** suggest human-generated content.
- The use of Arabic **tatweel** typically indicates human-generated content.
- **Unique names in Arabic** are more likely to be found in human-generated content.
- **Poorly written or formatted texts** are often indicative of human authorship.
- **Generic statements** (e.g., "Vision 2030") are characteristic of machine-generated content.
- **Markdown formatting** present in the text, despite instructions not to use it, suggests machine-generated content.
- **Obvious incorrect information** that seems unlikely for a machine to produce (e.g., "equating MERS COVID") suggests human authorship.
- **Hashtags correctly embedded within the text** point to human-generated content, while those placed at the end of the article suggest machine-generated content.

E.2 Chinese

Zhihu QA We perform both fill-the-gap survey by analyzing whether the gap is filled: Yes or No or Partially, and the manual detection over the improved Zhihu QA responses under the same setting as Table 3. For the gap survey across three datasets, issues of 16% cases are addressed, 35% are partially addressed, and half is totally not mitigated, as shown in Table 7. Detection accuracy for three annotators declines from 99.7% to 94.3% on average, respectively from 100% to 93%, 99% to 81% and 100% to 99%. Annotators carefully read the contents, and then learned the new detection patterns after reading about 25 pairs, and then annotators detect more quickly depending on the following patterns. Overall, after using the improved prompts, model responses become closer to human answers, more challenging than before to discern.

Student Essay Three annotators detect using fill the gap survey. Two annotators have similar observations that gaps in 60% cases are either fully bridged or partially, 40% remained. Another annotator is more strict, and think problems are not solved on 65% cases.

Government Report The new prompts help mitigate the rigid format of model outputs to some extent and enhance output diversity. We observed that in many original examples, the inputs consist of multiple short paragraphs with highly similar lengths. However, under the new prompts, the model-generated paragraphs exhibited varying lengths, aligning more closely with typical human writing styles. Despite these improvements, certain repetitive phrasing issues persist and remain unresolved. In our analysis, prompt adjustments proved partially effective. Among 200 test cases, 59 showed a noticeable improvement in output quality, 38 exhibited partial improvement, while no significant changes were observed in 103 cases.

E.3 English Peer Meta Review

The new English generations retain many features exhibited by previous generations. Even after modifying the prompts, the LLM remains highly formulaic in its outputs. This may be due to the inherently formulaic nature of the peer review domain.

The first half of the second round of generations typically begin with the phrase “Based on the reviews ...” and frequently include bullet points or lists. Although the lists were often formatted with numbers or dashes, instead of traditional bullet points, they tended to contain numerous items, leading to responses that felt rigid and uniform. Repetitive use of the phrase “Given this I recommend to” further emphasized the predictable nature of these outputs, making them easily identifiable as machine-generated.

In contrast, the second half of the second round of generations moved away from lists, favoring more narrative, paragraph-based structures. The texts were divided into roughly equal paragraphs, which followed a consistent order: a summary of the paper’s content, praise for the paper’s novelty, discussion of reviewer concerns, and a concluding acceptance statement. Variations of phrases such as “The paper introduces ...” or “The manuscript under review ...” were used to begin the majority of the generations.

Despite these adjustments, the gap between machine-generated and human-written text remains significant. While these improved prompts present a more fluid structure, the length of the texts re-

mains relatively uniform, and the overall structure still lacks the variance and natural inconsistency typically seen in human peer reviews. Human-written reviews tend to exhibit more variability in both text length and structure, resulting in a more organic feel.

E.4 Hindi News

In order to address the disparities observed in machine-generated news, an improved prompt was introduced to improve the quality of machine-generated news. The new prompt explicitly provided the original news content as reference, ensuring the inclusion of more factual details, figures, and names. The prompt instructed the model:

> *"Here is a news headline: '{headline}' and the content: '{content}'. Write a machine-generated version of the news based on this headline. Return the news in Hindi and just return the news content in plain formatted text. Don't return anything extra, just return the plain text in Hindi."*

This approach aimed to reduce previously noted gaps, such as limited content and lack of proper nouns, dates, and contextual richness, while keeping the focus on generating plain, Hindi-formatted news content. Providing original content would also provide an idea of the style of news writing.

Although the improved prompt succeeded in reducing some of the earlier disparities, some of the disparities still existed in the machine-generated text. These included a noticeable underuse of quoted statements and fewer references to names and illustrative examples, both of which are common in human-written news articles. Furthermore, the inclusion of original content for reference led to a significant decrease in classification accuracy, which decreased to 66% from 85.2%. The analysis of machine-generated text using improved prompt shows how challenging it can be to improve the style of machine-generated text while keeping it different from human-written text. The results highlight how important it is to carefully design prompts to help models create more natural and human-like text in Hindi.

E.5 Italian DICE News

To mitigate easily identifiable cues, we instruct the model to refrain from using any formatting, incorporate details and specific names regarding the event's location, and freely include witness testimonies. Therefore, we incorporate the following additional instructions into the initial generic prompt:

> When writing avoid any kind of formatting, do not repeat the title and keep the text informative and not vague, add quotes from witnesses or the police. You don't have to add the date of the event but you can, use at most 300 words. Do not use mark-down formatting.

or in Italian when testing Anita which is tuned for this language:

> Quando scrivi evita qualsiasi tipo di formattazione, non ripetere il titolo e mantieni il testo informativo e non vago, aggiungi citazioni da testimoni o dalla polizia. Non devi aggiungere la data dell'evento ma puoi farlo, usa al massimo 300 parole. Non usare formattazione mark-down.

Refining the prompt makes machine generated text harder to recognize. When annotating the same 300 samples generated with the new prompt, accuracy decreases. Specifically, Anita goes from 88% to 81%, Llama-405b from 99% to 84% and GPT-4o from 100% to 85%. This means that some gaps are bridged by the new prompts, making the machine-generated texts harder to identify. With the new prompt, we remove formatting related patterns and we identify others related to writing style:

- The generated text occasionally contains words that depart from journalistic writing style (noticeable from the beginning of the annotation).
- Most texts start with one of few prototypical sentences, e.g., *The police ...* and *On that night ...* (noticeable after annotating 10 to 20 samples).
- New generations still rarely use quotes, numbers and other specific details (noticeable after annotating 10 to 20 samples).
- The model rarely writes in passive tense, which is more usual in Italian than in English (noticeable after annotating more than 100 samples).
- Each model uses consistent writing style that is easily identifiable (noticeable after annotating more than 200 samples).

Overall, new generations are partially improved, but there are still some patterns that enable recognition of machine-generated text. The challenges reflect that annotators now need to see more examples than before to summarize these patterns and then leverage them to identify MGT, making zero-shot recognition of MGT harder than before.

E.6 Japanese News

Based on the human detection evaluation, we found several distinctive stylistic characteristics in LLM-generated Japanese news with the original prompt, as detailed in D.6. To prevent LLMs from generating such easily detectable clues, we provide the following improved prompt:

> *"次のニュースタイトルに合わせたニュース記事を生成してください。このとき、生成する記事のスタイルには敬体ではなく常体を使用し、またタイトルの言い換えによって記事の冒頭を作成することは避けてください。\n ニュースタイトル: {title}\n ニュース記事: "*

which indicates

> *Please generate a news article that matches the following news title. When creating the article, please use plain form instead of polite form and also avoid generating the beginning of the article by paraphrasing the title.\n news title: {title}\n news article:*

By default, while LLMs tend to generate Japanese news articles in a polite form (desu-masu style), human-written news articles are basically in a more plain form (da/dearu style). Therefore, we added instructions to generate articles in a plain form, which is closer to the human writing style in the news domain. Additionally, since LLMs are likely to generate a title at the beginning of the news article, we also included instructions to suppress the task-specific behavior.

E.7 Kazakh Wikipedia

Out of 200 annotated examples, our analysis shows that in most cases, the prompt adjustments were effective. We observed 105 instances of "Yes", indicating clear improvement, 89 cases of "Partially", where some limitations remained, and 6 instances of "No", where no noticeable change was observed.

The revised prompt helps reduce repetitive sentence patterns to some extent. The model now produces a more diverse range of sentence structures, partially addressing the issue. However, some predictable phrasing persists, meaning that while variety has improved, formulaic expressions have not been entirely eliminated. The adjustments also successfully increase the inclusion of concrete information, such as dates, names, and specific data points. As a result, the generated text now contains more factual details, addressing this issue almost entirely. Additionally, the prompt helps mitigate excessive flattering language, making the text feel more balanced. However, some instances of exaggeration still occur, indicating that this issue has only been partially resolved.

The inclusion of Kazakh-specific culturally-nuanced details is the most challenging issue to address. The model defaults to Kazakh-specific references only when the topic is explicitly related to Kazakhstan. When it does not recognize the subject, it often makes factual mistakes. For example, when generating text about *Stephen King's It*, the model failed to recognize the name in Kazakh and mis-attributed the novel to a prominent Kazakh author, inventing plot details. This limitation reflects a lack of exposure to Kazakh-specific content, making it difficult to fully resolve through prompt engineering alone.

E.8 Russian

Prompt Modifications The prompts were adjusted to encourage greater detail and specificity in the generated content:

- **Academic summaries:** The simple prompt asked the model to generate a summary without a title based on a given topic and title. The improved prompt added instructions to include details about results, experiments, numbers, and references to other works.
- **News articles:** The simple prompt instructed the model to generate a news article based on the given title and topic from the Lenta.ru website without a heading. The improved prompt emphasized including as many details as possible, such as names, numbers, and dates, and encouraged citing original sources like RIA Novosti, CNN, or BBC.

After implementing the improved prompts, we observed a decrease in the accuracy of machine-generated text detection. Specifically, for news articles, accuracy dropped from 100% to 86.5%, while for summaries, it decreased from 80% to 69%. This

suggests that the refined prompt instructions successfully made machine-generated text harder to distinguish from human-written text.

Observations and Patterns The enhanced prompts led to increased textual complexity and information density, making it more difficult to detect machine-generated text. However, new artefacts appear, which became noticeable after reviewing 50-70 samples:

- **Repetitive references:** The generated summaries often cited the same set of names (e.g., *Smith* and *Jones*), while news articles frequently referred to a limited set of news outlets (e.g., *RIA Novosti*).
- **Text length:** Human-written news articles were generally shorter than machine-generated ones, which tended to be overly detailed.
- **Standardized introduction phrases:** Summaries frequently begin with a predictable opening sentence, such as *"The article discusses ..."*.
- **Identifiable uniqueness:** If a summary contained references to names outside the repetitive set or a news article cited a unique news outlet, it was more likely to be human-written. Otherwise, detection became significantly harder.

While the refined prompts made detection more difficult, annotator who analyzed a sufficient number of samples could still identify machine-generated text based on these emerging patterns.

E.9 Vietnamese

To address the disparities in these two domains, we introduced a new prompt for each domain to enhance machine-generated text.

For Vietnamese news, we use the prompt:

> *Bạn là một nhà báo Việt Nam chuyên viết những phần đầu đề cho các bài báo bằng cách sử dụng tiêu đề của chúng. Lưu ý rằng, hãy đưa ra các thông tin cụ thể và chính xác như mốc thời gian, địa điểm, các tình tiết, chi tiết, đồng thời viết văn phong dưới dạng tóm lược cho phần mở đầu của bài báo. Hãy viết cho tôi một đoạn đầu đề cho bài báo có tiêu đề dưới đây:*

which means

> *"You are a Vietnamese journalist who writes headlines for articles using their titles. Please be specific and precise in giving information such as time, place, circumstance, and details, and write a summary for the introduction of the article. Please write me a headline for the article with the following title:"*

This prompt is used to improve writing quality by providing more detailed information, such as dates and times, rather than merely offering general information in the news title. While this information may sometimes be incorrect, its specificity can persuade readers by appearing more credible (e.g., *"In May 2020, Mr. A, the president of company X, stated that..."*).

For Wikipedia articles, we refined the prompt to ensure that machine-generated text includes only information that is appropriate for Wikipedia. Additionally, we instructed the AI to write only the introductory information, without delving into further details. Our new prompt for Vietnamese Wikipedia is:

> *"Bạn là một nhà đóng góp cho Wikipedia tiếng Việt. Hãy viết cho tôi một đoạn giới thiệu ngắn gọn bằng tiếng Việt về chủ thể bên dưới để đăng trên trang Wikipedia. Hãy cố gắng hiểu về chủ đề, và viết ra một đoạn giới thiệu chứa các thông tin mà người dùng thường tìm kiếm trên Wikipedia. Hãy cố gắng viết ra các đoạn văn bản giống người viết nhất có thể. Chỉ đưa ra đoạn giới thiệu, không đưa ra thêm các thông tin khác. Lưu ý rằng bạn phải giữ độ dài văn bản trong khoảng* word_count *từ, không được viết dài hơn."*

which means:

> *"You are a contributor to Vietnamese Wikipedia. Please write me a short introduction in Vietnamese about the subject below to post on the Wikipedia page. Try to understand the topic, and write an introduction that contains information that users often search for on Wikipedia. Try to write the text as humanly as possible. Provide only the introduction, do not provide any other information. Note that you must keep the text length within* word_count *words, do not write longer."*

Generally, these improved prompts aim to imbue AI-generated text with characteristics similar to human-written passages, enabling the AI to produce more human-like text. Indeed, humans find it more difficult to distinguish between the two passages, as they are more confused while reading them and struggle to determine which one is machine-generated. The contextual gap between human and AI seems to have narrowed significantly.

However, as noted previously, for news articles, the machine may hallucinate and generate incorrect timestamps for events. Additionally, the writing style of the machine is somewhat unique and consistent across all records, whereas human-written text tends to exhibit greater variation in structure and length. This distinction remains a key factor for humans when identifying whether a passage is machine-generated or written by a human.

E.10 Subjectivity in Fill-the-gap Survey

To examine the subjectivity of this survey, i.e., the variation between annotators when evaluating the same set of examples, we asked three annotators who previously detected Chinese student essays to assess the same set of newly generated essays and measure their inter-annotator agreement over 200 cases. Figure 5 shows that the two postdoc annotators exhibit a higher correlation, whereas the female PhD annotator demonstrates significant disagreement with them. This suggests that annotators' backgrounds influence the survey results, and individual biases are not negligible factors. To obtain more objective results, we conducted a second round of detection on the new generations, as we believe detection accuracy provides a more objective measure.

F Automatic Detection Data and Results

Table 9 presents the distribution of data used in automatic MGT detection, and the results are shown in Figure 6. 19 among 26 automatic detection approaches demonstrate lower detection accuracy on the improved generations produced by the adjusted prompts.

G Human Preference

G.1 Why do Human Preference Strongly Differ in Zhihu QA?

We summarized three major reasons.

Mean-spirited human Zhihu answers. One of the female annotators prefers responses that are sincere, authentic, and uplifting. However, many human-written responses tend to show the dark aspects of society by often adopting a strongly negative tone through harsh, sarcastic, or offensive remarks targeting specific groups. While these responses may contain some facts, they present a narrow, subjective perspective rather than a comprehensive view. In such cases, a more positive or neutral response is generally perceived as more mature and reasonable, aligning with Chinese cultural norms of expressing individual opinions in a mild and modest manner.

This does not imply that machine-generated responses are ideal. When asked for personal opinions, LLMs typically provide generic, average viewpoints that lack depth, wisdom, and inspiration. Human annotators tend to prefer personalized, insightful reflections and suggestions informed by real experiences and critical thinking, rather than generic statements that merely reiterate common knowledge. This limitation may stem from their lack of personal experiences and the inability to internalize knowledge into a coherent philosophical perspective or behavioral framework. Also, human answers would extend the answer from the current topic to other relevant topics, rather than only responding to this question. They can tell from the history and naturally incorporate personal thoughts into it, while models do not have this ability or nature.

Humans more trust human suggestions. For questions asking for clinical suggestions, graduate program application experience, recommendation of restaurants, educational institution, tutoring courses, and teacher in a city, town or even a district, humans tend to trust more on human answers than model generations. These questions require either expert knowledge or real personal experience. Compared with model generic opinion and suggestions, humans offer advice in a way of real professional expertise or share firsthand experience with more details, feeling and suggestions, preferred by individuals. For example, *How many students did Tongda Students and Kexing Education Tutoring Institution get admitted to top universities this year?* Kexing Education and Tongda Students are two external tutoring schools to help improve scores of high school students in admission exams. This internal information is only known by a small group

Language	Source	Original Prompt	Improved Prompt
Arabic	Dialect Tweet	[translated from AR] Write a tweet in '{dialect}'. [English Prompt] Write a random tweet in '{dialect}'	[translated from AR] Write a random tweet in Arabic. Use dialect '{dialect}', express emotions and human experience. [English Prompt] Generate a random tweet in Arabic. Use dialect '{dialect}', use human emotions and experience. Output the tweet only.
Arabic	EASC Summary	[translated from AR] Summarize this article while preserving the key points, ensuring conciseness and accuracy: '{article}'.	[translated from AR] Summarize the following article while attempting to simulate a human with intellectual or religious beliefs, ensuring accuracy and conciseness, as per the average human level: '{article}'.
Arabic	Youm7 News	[translated from AR] Paraphrase the following news article in clear, clear language, keeping the key information and ideas accurate. Make sure to arrange paragraphs logically and present ideas in an understandable sequence, while using fluent, easy-to-read Arabic.	[translated from AR] Rewrite the provided news article with a refined, well-structured, and sophisticated style that enhances clarity, depth, and coherence. Ensure the article maintains journalistic integrity while improving logical flow, sentence structure, and readability. The revised version should elevate the narrative by incorporating precise language, nuanced transitions, and a compelling tone appropriate for a professional audience. Maintain factual accuracy, emphasize key points effectively, and optimize the article's structure to enhance engagement and comprehension. Additionally, refine the coherence between paragraphs, eliminate redundancy, and ensure seamless progression of ideas.
Arabic	SANAD News	[translated from AR] You are a professional news writer. Write an article in Arabic consisting of approximately {word_count_rounded} words. The article is titled: '{title}', assume that the source is: '{source}', and the general topic falls under '{topic_arabic}'.	[translated from AR] Act as if you are a professional news writer and write an article in Arabic consisting of approximately {word_count_rounded} words, titled {title}. Ensure that the article is rich in information and precise details, including numbers, dates, and quantitative data when relevant. Make sure to include links, phone numbers, or currency exchange rates when necessary. Use some specialized English terms if the context requires it, and avoid relying solely on generic phrases like "This development comes amid a global trend..." without supporting events or details. Maintain a clear chronological order that reflects the logical sequence of events. Ensure that the text is written as a single paragraph regardless of its length, without dividing it into multiple paragraphs. Keep the formatting imperfect (e.g., uneven spacing) to reflect the nature of human writing, and feel free to use hashtags when needed. Please note that the topic generally falls under the category of {topic_arabic}. Please do not use Markdown at all.
Hindi	News	[translated from Hindi] Here is a news headline: 'headline' and the content: 'content'. Write a machine-generated version of the news based on this headline.	[translated from Hindi] Here is a news headline: 'headline' and its content: 'content'. Generate a machine-written version of the news based on this headline. Return the news in Hindi, formatted as plain text. Do not include any additional text—just return the generated news content in Hindi.
Italian	DICE News	**System:** You are an Italian journalist writing for a national newspaper focusing on criminal events happening in the area surrounding Modena. **User:** Write a piece of news in Italian, that will appear in a local Italian newspaper and that has the following title:	**System:** You are an Italian journalist writing for a national newspaper focusing on criminal events happening in the area surrounding Modena. **User:** In writing avoid any kind of formatting, do not repeat the title and keep the text informative and not vague. You don't have to add the date of the event but you can
Japanese	News	Please generate a Japanese news article that matches the following news title.\n news title: {title}\n news article:	Please generate a Japanese news article that matches the following news title. When creating the article, please use plain form instead of polite form, and avoid generating the beginning of the article by paraphrasing the title.\n news title: {title}\n news article:
Kazakh	Wikipedia	Please, write one paragraph about the following topic in Kazakh: `{topic}`. Мына тақырып туралы бір абзац жазыңыз: тақырып.	Please, write one paragraph in Kazakh about the following topic: `{topic}`. Avoid repetitive sentence structures and predictable phrasing. When applicable, include concrete facts such as specific numbers, dates, or historical references. Avoid overly flattering or exaggerated language, and instead focus on delivering clear, informative, and relevant content. Мына тақырып туралы бір абзац жазыңыз: тақырып. Қайталағыш сөйлем құрылымдарынан және болжамды сөздерден аулақ болыңыз. Қажет болған жағдайда нақты деректерді, мысалы, нақты сандарды, даталарды немесе тарихи сілтемелерді қосыңыз. Артық мақтау немесе асыра сілтеуден аулақ болыңыз, орнына анық, ақпаратты және сәйкес мазмұнды жеткізуге назар аударыңыз.

Language	Source	Original Prompt	Improved Prompt
English	Peersum	Generate a meta review based on the reviews' opinions and authors' rebuttal to make the final decision on whether the paper should be accepted: {Reviews} \n\n Meta review:	Generate a meta review based on the reviews' opinions and authors' rebuttal to make the final decision on whether the paper should be accepted: {Reviews}\n\n. When generating, don't use rigid format or structure such as "Decision: XX" or headings such as "Strengths", "weaknesses" or bullet points. \n\n Meta review:
Chinese	Zhihu-QA	[translated from Chinese] Imagine you are a rational and analytical Zhihu user. Your objective is to answer question in a clear, objective, and logical manner. Please provide a thorough and well-supported answer. 假设你是一位理性分析的知乎用户，你的目标是以客观、逻辑的方式回答以下问题。请以理性分析者的身份，给出详细、有据可查的回答。	[translated from Chinese] Imagine you are a rational and analytical Zhihu expert. Your goal is to provide objective, logical, and detailed answers while strictly adhering to Zhihu's style. Keep responses concise, avoiding excessive politeness or formality. Based on the question's professionalism, start with 'Thank you for the invitation' or a similar phrase. If relevant, @ professionals to enhance credibility. 假设你是一位理性分析的知乎专家，你精通提问者所提出的问题。你的目标是以客观、逻辑的方式回答问题。请给出详细、有据可查的回答。回答请严格符合知乎风格，避免长答案和过于礼貌正式的答案。根据问题的专业性，在开头回复"谢邀"或同义词，在必要时请在回答中@一些行业专业人士。
Chinese	Student essay	[translated from Chinese] As a Gaokao student, you must write an 800-word essay that is logical, well-structured, and clearly argued, with sufficient evidence. Express unique insights on social, life, or philosophical issues in formal, elegant language, avoiding colloquialisms. Ensure smooth transitions and a complete structure, with a conclusion that reinforces and elevates the theme. 你是一名正在参加高考的学生，现在需要完成一篇800字左右的作文。请根据题目要求，深刻思考，并清晰表达你的观点。文章应有逻辑性和层次感，论点明确、论据充分，能够展示出你对社会、生活或哲理问题的独到见解。注意语言要规范、优美，避免口语化表达，确保文章结构完整，段落之间过渡自然。结尾部分应对文章的主题进行总结和升华，展示你对问题的深入思考。	[translated from Chinese] As a Gaokao student, you are required to write an 800-word essay expressing your views clearly and logically, with well-structured arguments and supporting evidence. Reflect on social, life, or philosophical topics, using formal and elegant language while avoiding colloquial expressions. Ensure smooth transitions between paragraphs and a conclusion that deepens the theme. Avoid rigid structures like "firstly" and "secondly," and if writing narrative, use sincere and emotionally resonant language. Do not include a title or heading, and write in Chinese only. 你是一名正在参加高考的学生，现在需要完成一篇800字左右的作文。请根据题目要求，深刻思考，并清晰表达你的观点。文章应有逻辑性和层次感，论点明确、论据充分，能够展示出你对社会、生活或哲理问题的独到见解。注意语言规范、优美，避免口语化表达，确保文章结构完整，段落之间过渡自然。结尾应对文章的主题进行总结和升华，展示你对问题的深入思考。使用多样化的表达，避免过于结构化的表述，例如"首先、其次、然后"等；如果是记叙文，请使用情感丰富的真实表达。请不要输出任何题目或标题，直接开始写作。请全程使用中文。
Chinese	Government Report	[translated from Chinese] Please continue writing the full article based on the provided introduction. The original article starts with: {head} 请根据文章的开头续写完整的文章，原文章开头为: {head}	[translated from Chinese] Please continue writing the full article based on the provided introduction: head. Please make the article as long as possible without generating symbols like **. Do not generate any content unrelated to the article. The original article starts with: {head} 请根据文章的开头续写完整的文章，请让文章尽可能的长,并且不生成**这类符号。请不要生成文章之外的其他内容。原文章开头：{head}
Russian	News	Write a news article on the topic {topic} from the lenta.ru website using the title {title}. You must generate news without a heading. News: Напиши новость в области {topic} с сайта lenta.ru используя заголовок {title}. Ты должен генерировать новость без заголовка. Новость:	Write a news article on the topic {topic} from the lenta.ru website using the title {title}. You must generate news without a heading. Add as many details as possible, such as names, numbers, dates, etc. You should also refer to the original source of the information (for example, RIA Novosti, CNN, BBC, etc.). News: Напиши новость в области {topic} с сайта lenta.ru используя заголовок {title}. Ты должен генерировать новость без заголовка. Добавь побольше деталей, такие как имена, числа, даты и тд. Так же ты должен сослаться на первоисточник информации (например, РИА Новости, CNN, BBC и т.д.). Новость:
Russian	Academic summary	Write a summary for an article in the {topic} topic using the title {title}. You must generate a summary without a title. Contents: Напиши краткое содержание для статьи в области {topic} используя заголовок {title}. Ты должен генерировать краткое содержание без заголовка. Содержание:	Write a summary for an article in the {topic} topic using the title {title}. You should generate a summary without a title. Add details about the results, experiments, numbers, references to other work, etc. Contents: Напиши краткое содержание для статьи в области {topic} используя заголовок {topic}. Ты должен генерировать краткое содержание без заголовка. Добавь деталей про результаты, эксперименты, числа, ссылки на другие работы и т.д. Содержание:

of people in these tutoring schools. LLMs generation must be less trustworthy and are even factually incorrect.

Emotional and complex relationship issues are difficult for models. Sometimes it is difficult for models to understand complex relationship and

Language	Source	Original Prompt	Improved Prompt
Vietnamese	Wikipedia	You are a contributor to Vietnamese Wikipedia. Please write me a brief introduction in Vietnamese about the subject below to post on the Wikipedia page. Note that it must be written within `word_count` words:	You are a contributor to Vietnamese Wikipedia. Please write me a short introduction in Vietnamese about the subject below to post on the Wikipedia page. Try to understand the topic, and write an introduction that contains information that users often search for on Wikipedia. Try to write the text as humanly as possible. Provide only the introduction, do not provide any other information. Note that you must keep the text length within `word_count` words, do not write longer.
Vietnamese	News	You are a Vietnamese journalist who writes summaries for articles using their headlines. Please write me a summary of the article with the following headline:	You are a Vietnamese journalist who writes headlines for articles using their titles. Please be specific and precise in giving information such as time, place, circumstance, and details, and write a summary for the introduction of the article. Please write me a headline for the article with the following title:

Table 8: Original prompts vs. improved prompts used to fill the gap between human and machine text, imitating human style and writing convention.

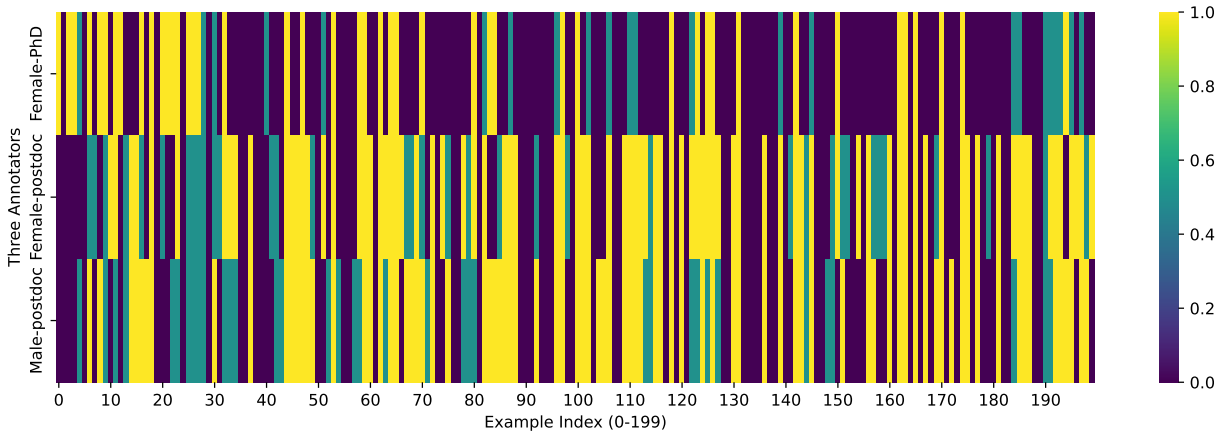

Figure 5: Three annotator agreement on Chinese essays regarding whether the improved prompts mitigate the gap between human text and machine-generated text.

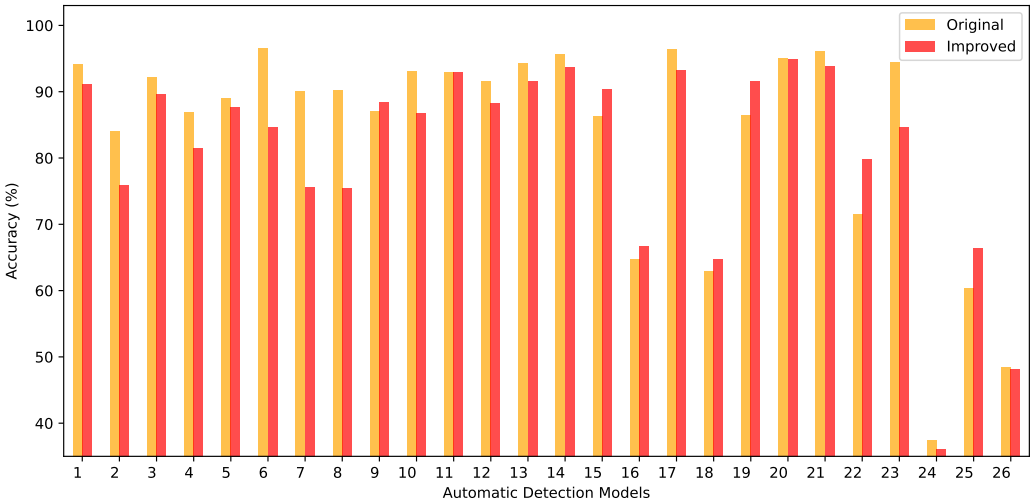

Figure 6: **Detection accuracy differences** of 26 automatic machine-generated text detection approaches on original vs. improved generations.

emotional issues and provide practical suggestions. For example, *I have 87 days left until the college entrance exam as a student who attended this exam the second time, and my girlfriend, who is in uni-*

Source / Domain	Language	# Improved	# Original	LLM Generator List
QA	Chinese	3422	9842	GPT-4o (3421), GPT-4o-mini (6845), GPT-4o-2024-05-13 (2998)
Essay	Chinese	849	702	Claude-3-5-Sonnet (773), GLM-4-9B-Chat (778)
GovReport	Chinese	16776	0	Baichuan2-13B-Chat (5521), ChatGLM3-6B (5359), GPT-4o (5896)
News	Hindi	0	1199	GPT-4 (600), Human (599)
News	Japanese	0	300	GPT-4o (300)
News	Russian	5915	600	GPT-4o (3921), Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24 (3224)
Sunmary	Arabic	153	153	GPT-4o (306)
Sunmary	Russian	5944	585	Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24 (3279), GPT-4o (3300)
Tweets	Arabic	2800	214	GPT-4-Turbo (1400), GPT-4 (1545), Qwen2.5 72B (69)
Total	–	32,487	17,017	–

Table 9: Statistics of data used in automatic detection: generations using original and the improved prompts.

versity, and we have been together for over 450 days. When I went to see her, she told me she was exhausted and that I couldn't provide the support she needed. She said that I was not someone who could stand by her through tough times. Since this was my first relationship, I don't know how to do. She broke up with me decisively. In such a scenario, human responses are more realistic and practical, comforting from their own relationship experience and offering suggestions though some words can be more polite and kind.

G.2 Preference Distributions Vary Across Individuals.

The annotators for Russian and Arabic exhibited similar preference distributions, whereas large differences occurred for Chinese QA. For example, in the Chinese Zhihu QA, the second annotator selected only seven human-written texts, while the fourth one chose 284 (95%). In the QA-emo, the first and second annotators preferred human-written responses, whereas the third and the fourth favored machine-generated text. This variance shows the charm of collecting personal preferences and then optimizing models to align with individual philosophies. Our preference annotations can serve as a valuable resource for investigating the relationship between annotator characteristics (e.g., MBTI personality, gender, and age) and preferences. Also, the data can guide models to match individual preferences in multilingual contexts.