SPACSEC'2026 · SAINT PETERSBURG · 26–28 OCTOBER 2026
International Conference on Cyber Warfare, Security & Space Computing · ITMO University



Investigating the Robustness of a Classifier for Texts Generated by LLMs to Automatic Paraphrasing

Nikita Kurganov

Bauman Moscow State Technical University (BMSTU) · kurganov.n@bmstu.ru

BACKGROUND

Motivation

Modern LLMs can deeply rewrite text, which makes paraphrasing a natural way to evade AI-text detectors.

01

Deep paraphrasing

LLMs rewrite existing text while preserving its meaning, argument structure and key facts.

02

Detector evasion

Automatic paraphrasing can be used to bypass systems that detect LLM-generated text.

03

Russian is understudied

Most studies focus on English and ignore Russian morphology, syntax and style. RuATD 2022 did not test robustness to LLM paraphrasing.

BACKGROUND

Aim and Tasks

Is a Russian-language LLM-text detector robust when the text is rewritten by another LLM?

AIM

Study the robustness of a previously developed classifier of Russian-language LLM-generated texts, based on the sbert_large_nlu_ru encoder, to automatic paraphrasing by large language models.

PRACTICAL VALUE

The results support the development of detectors for LLM-generated and partially modified texts.

TASKS

1. Formalize text modification levels from 0% to 100%
2. Measure how classification changes when texts are paraphrased by GPT-4.1, Llama-3.3-70B and t-pro-it-1-2
3. Analyze classification errors and the statistical significance of the changes

DATA & MODEL

Training and Evaluation Data

The classifier was trained on one corpus and tested on four independent Russian-language datasets.

TRAINING · RUTEXTLLMDETECT

601K

texts: ~300K human-written + ~300K generated by

`GPT-4o` `t-pro-1.0` `Llama-3.3-70B` `ruGPT3-Large` `mT5-Large` `Facebook-XGLM`

Human texts: news, scientific articles and web content published no later than 2019.

INDEPENDENT EVALUATION

52K

texts from four public Russian-language datasets

Dataset	Total	LLM	Human
M4	12,000	6,000	6,000
MULTITuDE	9,074	7,775	1,299
RuATD-2022-BI	21,511	10,755	10,756
open-rus-mgtd	10,000	5,000	5,000

DATA & MODEL

Classifier Architecture

Hybrid model: a pre-trained transformer encoder, a bidirectional LSTM and fully connected layers.

01
Tokenization
input_ids, attention_mask; up to 512 tokens

→

02
Encoder
`sbert_large_nlu_ru`
pre-trained transformer: a feature vector for every token

→

03
BiLSTM
Dropout · bidirectional LSTM, h → 2h · Dropout

→

04
Residual block
FC 2h → h · BatchNorm · ReLU · FC h → h · BatchNorm · skip connection + LayerNorm

→

05
Classifier head
Dropout · FC · GELU · Dropout · FC h → 2: class probabilities

32
texts per training batch

512
max input length, tokens

2
output classes: human, LLM

0.5
decision threshold: LLM if P(LLM) ≥ 0.5

A100
NVIDIA Tesla A100 GPU, 80 GB

DATA & MODEL

Baseline Quality

Combined independent set of about 52K texts; the best overall result among previously compared architectures.

ACCURACY	PRECISION	RECALL	F1-SCORE
81.46%	85.46%	80.72%	83.02%

ACCURACY BY DATASET



Recall for the LLM class, 80.72%, is the 0% baseline for the paraphrasing experiments.

METHOD

Types of Text Transformation

Paraphrasing is controlled through four kinds of change applied to formal units of the text.



Lexical

`unit: significant words`

A significant word is replaced with a synonym, a phrase with a description; significant words are deleted or added without changing the meaning.



Morphological

`unit: word forms`

The grammatical form changes, the lexical meaning does not: verb tense, number, case; participial phrases are replaced.



Syntactic

`unit: sentences`

Sentence structure changes, the meaning stays: parts are reordered, a complex sentence is split, simple sentences are merged.



Compositional

`unit: paragraphs, blocks`

Sentences within a paragraph or whole semantic blocks are reordered; paragraphs are split and merged.

Significant words carry independent lexical meaning: nouns, adjectives, verbs, adverbs, participles, numerals. **Function words** (prepositions, conjunctions, particles, interjections) serve a syntactic role.

METHOD

Modification Level Scale

Ten cumulative levels set by prompt requirements, not a measured share of changed tokens.

Level	Lexical	Morphological	Syntactic	Compositional
10%	≤10% of significant words	≤5% of words	≤10%: one transformation	none
20%	10–20%	≤10%	≤20%: complex ↔ simple	none
30%	20–30%, incl. phrases	≤15%	≤30% change structure	reorder ≤10% of sentences
40%	30–40%	≤20%	≤40% modified	reorder within a paragraph
50%	40–50%, incl. constructions	≤25%	≥50% modified	reorder semantic fragments
60%	50–60%	≤30%	≥60% fully restructured	merge and split paragraphs
70%	60–70%, incl. speech patterns	≤35%	≥70% change structure type	reorder semantic blocks
80%	70–80%	≤40%	80% formed anew	full recomposition allowed
90%	80–90%, incl. constructions	≤45%	90% formed anew	new composition formed
100%	complete lexical replacement	any changes	all sentences formed anew	composition fully replaced

METHOD

Experimental Setup and Metrics

Each source text is paraphrased at every level by three LLMs and then classified.

01

Source texts

Independent set of about 52K texts, LLM-generated and human-written

→

02

Paraphrasing

`GPT-4.1`
`Llama-3.3-70B`
`t-pro-it-1-2`

levels 10–100%

→

03

Classification

sbert_large_nlu_ru classifier: LLM if P(LLM) is 0.5 or higher, otherwise human

→

04

Evaluation

Each level is compared with the original texts at 0%

Prompt: source text + target level + requirements for lexical, morphological, syntactic and compositional changes.

LLM-GENERATED TEXTS

Accuracy: the share of texts classified as LLM at each level

HUMAN-WRITTEN TEXTS

Mean and median probability of the human class

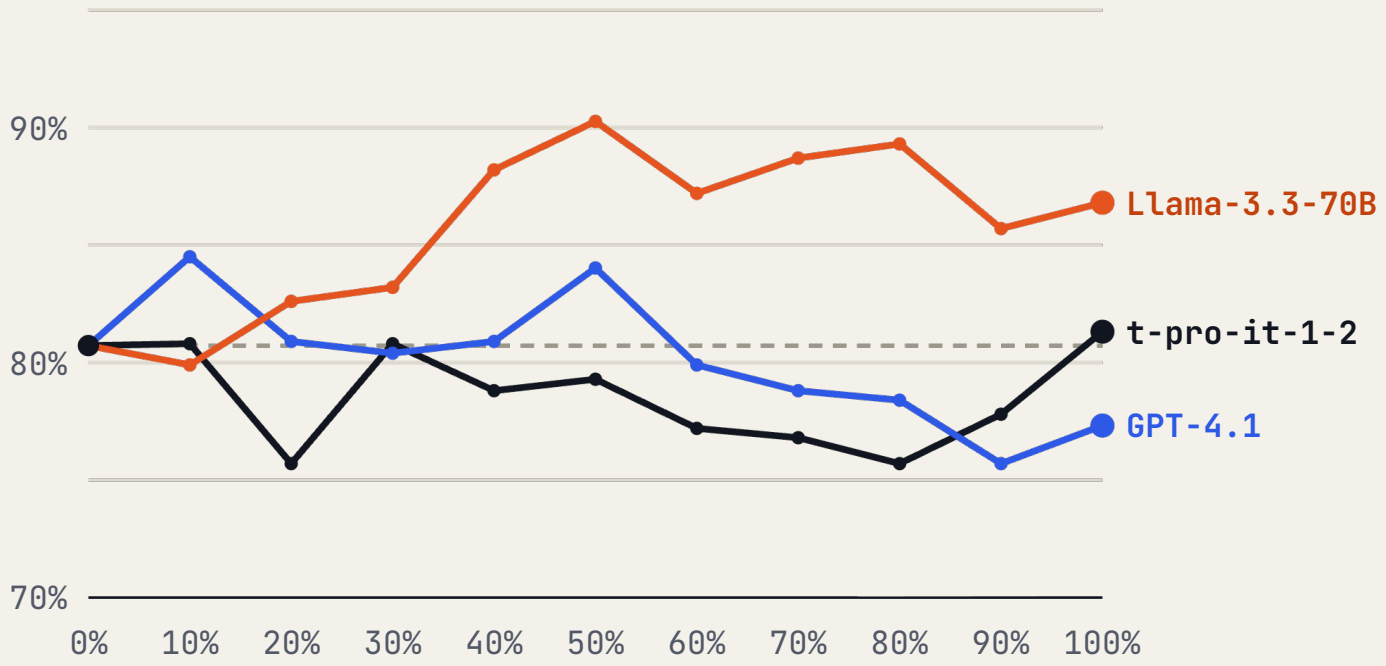
SIGNIFICANCE

McNemar test for paired decisions: levels 50% and 100% vs 0%, threshold p = 0.05

RESULTS

LLM-Generated Texts

Accuracy stays close to the 80.72% baseline at every modification level.



– – – no paraphrasing (level 0%): 80.72%

x: modification level · y: Accuracy

AT 100% MODIFICATION

Llama-3.3-70B

86.80% +6.08 pp

t-pro-it-1-2

81.31% +0.60 pp, n.s.

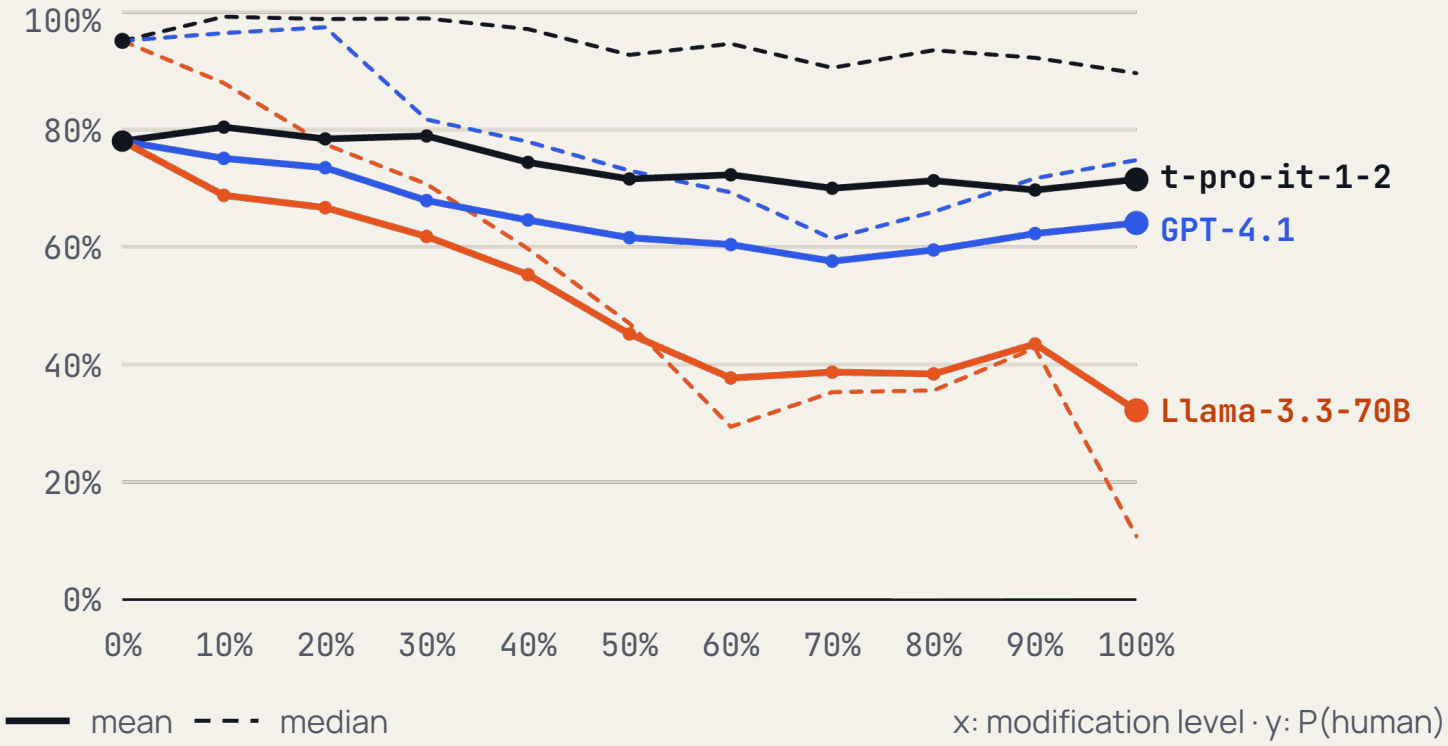
GPT-4.1

77.32% −3.40 pp

RESULTS

Human-Written Texts

As LLM paraphrasing grows, the probability of the human class decreases, most of all for Llama-3.3-70B.



P(HUMAN) AT 100%

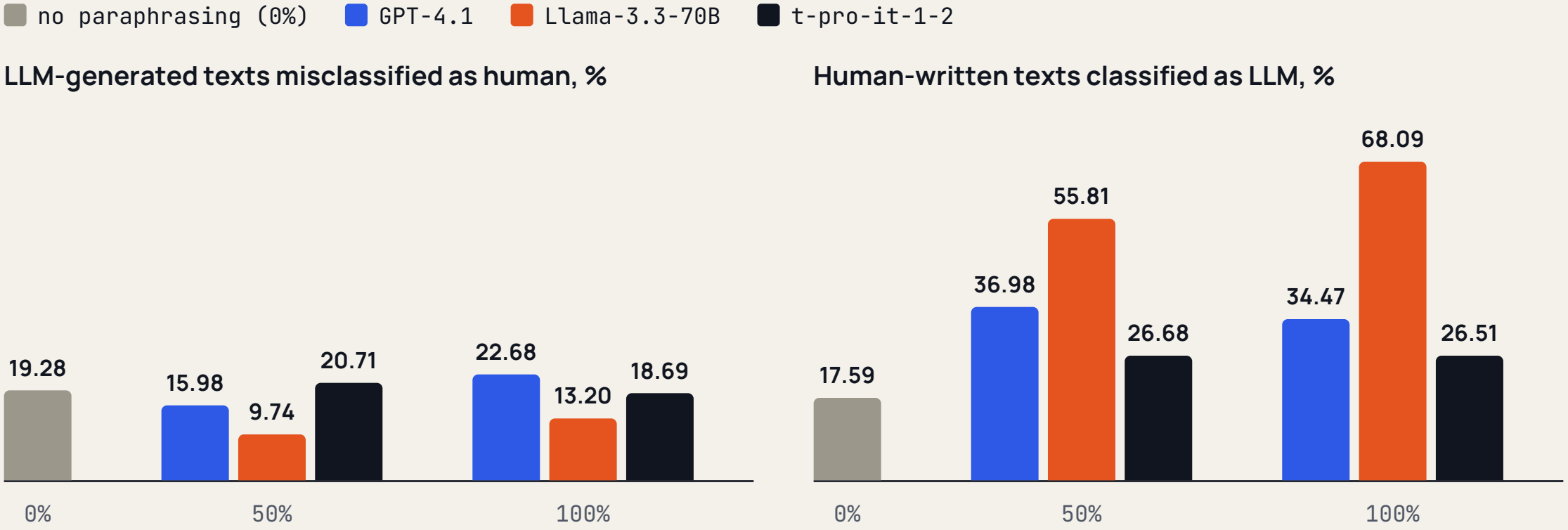
Model	Mean	Median
no paraphrasing	78.02%	95.09%
t-pro-it-1-2	71.50%	89.60%
GPT-4.1	64.10%	74.80%
Llama-3.3-70B	32.20%	10.80%

A lower P(human) is not necessarily an error: after paraphrasing, part of the text really is written by an LLM.

RESULTS

Classification Error Analysis

Errors on LLM texts do not grow systematically, while human texts drift toward the LLM class.



RESULTS

Statistical Significance

McNemar test vs the 0% level: every change is significant except t-pro-it-1-2 on LLM texts at 100%.

	LLM-generated · Δ Accuracy		Human-written · Δ share classified as LLM	
Model	**50%**	**100%**	**50%**	**100%**
GPT-4.1	+3.30 / < 0.001	−3.40 / < 0.001	+19.39 / < 0.001	+16.88 / < 0.001
Llama-3.3-70B	+9.54 / < 0.001	+6.08 / < 0.001	+38.22 / < 0.001	+50.50 / < 0.001
t-pro-it-1-2	−1.43 / < 0.001	+0.60 / 0.098 n.s.	+9.09 / < 0.001	+8.92 / < 0.001

Cells: change vs level 0% in percentage points / p-value of the McNemar test on paired decisions; n.s. = not significant at p = 0.05.

LLM-GENERATED TEXTS

The only significant drop in accuracy is GPT-4.1 at 100% (−3.40 pp); Llama-3.3-70B raises it by 6.08 pp.

HUMAN-WRITTEN TEXTS

The shift toward the LLM class is significant for all models ($p < 0.001$), up to +50.50 pp for Llama-3.3-70B.

KEY RESULT

Paraphrasing with GPT-4.1, Llama-3.3-70B and t-pro-it-1-2 does not provide a stable bypass of the classifier, even at 100% modification.

no paraphrasing	GPT-4.1	Llama-3.3-70B	t-pro-it-1-2
80.72%	77.32%	86.80%	81.31%

Accuracy on originally LLM-generated texts at the maximum modification level

CONCLUSION

Conclusions

Three findings, the limits of this study and the next steps.

1 The classifier is robust to the studied paraphrasing attacks: at 100% modification, accuracy on LLM texts is 77.32% / 86.80% / 81.31% vs 80.72% without paraphrasing.

2 For human-written texts, decisions shift toward the LLM class as paraphrasing grows: 34.47% / 68.09% / 26.51% are classified as LLM at 100%.

3 The outcome depends on the paraphrasing model, not only on the modification level.

Values: GPT-4.1 / Llama-3.3-70B / t-pro-it-1-2

LIMITATIONS

Three paraphrasing models and one classifier

Levels are set by transformation requirements, not by a measured share of changed elements

FUTURE WORK

More paraphrasing models and classifiers

Methods to estimate how much an LLM was involved in creating or modifying a text



Thank you!

Questions are welcome.

Speaker
Nikita Kurganov, BMSTU

E-mail
kurganov.n@bmstu.ru

ORCID
0009-0002-3656-8637