Russian-Language Text Detectors: A Three-Text Case with an Inverted Verdict

Author: Anton Grigoryants, Institute of Sociology and AI Safety (ISAIS) — research project, in the process of establishment. **Version:** 2 (31 October 2026; first version — 26 October 2026). **Type:** research note (case study), not a benchmark.

Abstract

We ran a small controlled case with three Russian-language texts through two commercial AI-text detectors on the same day, with a control example. The pure machine-written control was flagged by both services (100% AI). The two realistic texts produced an inversion: a business letter drafted by a model and rewritten by a human in its final paragraph was scored 93% AI by Pangram 4.1 and 0% (human) by a Russian service; a different business letter that was entirely machine-written, pasted from a chat window and signed by a person — with two identical typos — was scored “Possibly human, 100%” by Pangram. Two detectors disagreed by 93 points on the same input, in opposite directions, and the most confidently “human” verdict went to the most machine-generated text in the set.

Why this case exists

Detectors are increasingly used as evidence: in publishing, in education, in hiring. Russian is one of the languages major vendors list as supported (Pangram states 20+ languages and, for each, “99%+ accuracy”), while published Russian-language accuracy figures are scarce. A case that is fully documented — texts, verdicts, timestamps, limitations — is more useful than another opinion about whether detectors “work”.

Method

Three texts, Russian, one session, one day (26 October 2026).

1. **Control.** A letter written entirely by a large language model. The letter was copied in full, but only part of it reached the vendor’s interface: the text actually submitted for scoring was a **73-word excerpt**, uniform in tone, with the standard business formulas (“It is important to note”) and opening and closing salutations, and no human editing. Pangram counted the excerpt and flagged it as a short text.
2. **Machine-drafted, human-finished.** A business letter drafted by a model (≈185 words) and rewritten by a human in its final paragraph

— a short, informal paragraph with a personal remark. The body remained machine text.

3. **Machine-written, human-pasted.** A business letter written by a model in a chat interface, pasted into an email client and signed by a person (≈99 words). It contains bureaucratic constructions, a single 60-word sentence with two parentheticals, non-standard line wrapping from the chat window, and the same spelling error twice (“соответсвенно”).

Detectors. Pangram 4.1 (web application, trial mode, 26 October 2026); a Russian commercial service with an AI-text detector mode (Russian-language vendor). Scores and labels were recorded as displayed by each interface. The texts are anonymised: no personal names, no company names; the business letters were read by their addressees before this note, and the content is not reproduced here.

Results

Text	Pangram 4.1	Russian service
Control, pure machine (73-word excerpt)	“Possibly AI”, 100%	100% AI
Machine-drafted, human final paragraph (≈185 words)	**“AI Detected”, 93%** — with the note *“mostly AI-generated, human-written ending”*	**0%, “human”**
Machine-written, pasted and signed, two typos (≈99 words)	**“Possibly human”, 100%**	0%, “human”

Two observations.

First, the disagreement. On text 2 the services differ by 93 percentage points: 93% AI versus 0% AI. These are not rounding differences; they are opposite verdicts on the same input.

Second, the inversion. Text 3 is the most obviously machine-generated of the three by any structural reading: one giant sentence, stacked parentheticals, bureaucratic filler, and a repeated typo. A human reader in the same language sees the machine immediately. Pangram returned “Possibly human, 100%” (with its own warning that short texts are classified with less certainty). Meanwhile text 2, where a person actually rewrote the ending, was flagged as AI — although Pangram’s segment-level note correctly identified the human-written ending.

Discussion

The pattern is consistent with what independent evaluations report about what these systems measure: not authorship, but the statistical smoothness

of the text. A smooth, neatly edited machine draft is “AI”; a bureaucratic text with typos and irregular line breaks is “human”, even when it was generated wholesale. The “human-written ending” feature is useful evidence that the models can localise human editing — and that global percentages hide exactly the information a reader needs.

Three practical implications:

1. **Percentages are not comparable across vendors.** A single number without the service name, version, language and text length carries almost no information. Reporting “93% AI” as if it were an instrument reading is misleading.
2. **Russian remains under-tested.** Independent English-language results for the leading detector are strong (an NBER working paper from 2025 reports a false-positive rate at or below 0.5% for Pangram in English). No comparable independent Russian data exists. Our case is one data point, not a benchmark — but it sits in that gap.
3. **Human review remains the only working control.** Where the consequence is a reputation, a grade or a contract, the decision should rest on the text and its author, not on a percentage.

Limitations

One session, three texts, trial mode, one English-centric vendor and one Russian vendor. The case is not a benchmark: no corpus sampling, no repeated runs, no segment-level analysis beyond the vendor’s own labels. Two of the three texts were short enough for Pangram to flag reduced reliability: the control was a 73-word excerpt of a longer machine-written letter, and the pasted letter was 99 words. Only text 2 (185 words) sat above the length the vendor treats as reliable. Russian-language support in the leading detector is vendor-claimed, not independently validated. No claim is made about the vendors’ overall accuracy.

Correction

31 October 2026. Version 1 of this note described the control sample as a “≈180-word letter”. That figure was the length of the letter the model produced in full; what actually reached the vendor’s interface — and what both services scored — was a 73-word excerpt of it. The table and the Limitations section above are corrected accordingly, and the heading now carries the version date. The verdicts, the percentages and the conclusion do not change: on the 185-word letter the two services differ by 93 percentage points, and the most confident “human” verdict went to a text that was machine-written throughout.

The correction is published rather than made silently: a case study is only worth as much as its record of what was actually submitted.

Sources

1. Pangram — multilingual support statement: https://www.pangram.com/solutions/multilingual
2. NBER Working Paper 34223, “Artificial Writing and Automated Detection” (Jabarian & Imas, 8 September 2025): https://www.nber.org/system/files/working_papers/w34223/w34223.pdf
3. Our own measurements, 26 October 2026 (screenshots with verdicts, labels and word counts retained).

How to cite

Grigoryants, A. (2026). *Russian-Language Text Detectors: A Three-Text Case with an Inverted Verdict*. ISAIS research note, 26 October 2026. https://isai.su/notes/russian-detectors-2026-10-26.pdf