

PAN 2026

Lab on Digital Text Forensics and Stylometry



pan.webis.de



pan@webis.de

**Overview: “Voight-Kampff”
Generative AI / LLM Detection**

Janek Bevendorff¹

Rohit Raj Gunti²

Jussi Karlgren³

Maik Fröbe^{1,4}

Martin Potthast^{5,6,7}

Benno Stein¹

¹ Bauhaus-Universität Weimar

² University of Tennessee

³ University of Helsinki

⁴ Friedrich-Schiller-Universität Jena

⁵ University of Kassel

⁶ hessian.AI

⁷ ScaDS.AI

“Voight-Kampff”^{*} Generative AI / LLM Detection

Given a (potentially obfuscated) text, decide whether it was written by a human or an AI.

^{*} From the 1982 science fiction film *Blade Runner*. The Voight-Kampff is a polygraph-like machine used by blade runners to determine whether an individual is a replicant. [\[Wikipedia\]](#)

“Voight-Kampff”^{*} Generative AI / LLM Detection

Given a (potentially obfuscated) text, decide whether it was written by a human or an AI.

Organized as builder-breaker evaluation between two labs:

- ❑ **Builder** (137 Submissions / 34 Teams)

PAN participants develop classification algorithms to discriminate human authors and LLMs.

- ❑ **Breaker** (16 Submissions / 5 Teams)

ELOQUENT participants provide evaluation data to attack the *PAN* participants' classifiers.

^{*} From the 1982 science fiction film *Blade Runner*. The Voight-Kampff is a polygraph-like machine used by blade runners to determine whether an individual is a replicant. [\[Wikipedia\]](#)

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset Creation (Training)

Dataset: “PAN AI Obfuscated 2025” (Training)

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset Creation (Training)

Dataset: “PAN AI Obfuscated 2025” (Training)

- ❑ *Human (1)*: 7,300 19th-century novels (500–700 words).
Scraped from Project Gutenberg.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset Creation (Training)

Dataset: “PAN AI Obfuscated 2025” (Training)

- ❑ *Human (1)*: 7,300 19th-century novels (500–700 words).
Scraped from Project Gutenberg.
- ❑ *Human (2)*: 931 essays.
Brennan-Greenstadt (Brennan et. al, 2012) and Riddell-Juola (Wang et al., 2021) corpora.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset Creation (Training)

Dataset: “PAN AI Obfuscated 2025” (Training)

- ❑ *Human (1)*: 7,300 19th-century novels (500–700 words).
Scraped from Project Gutenberg.
- ❑ *Human (2)*: 931 essays.
Brennan-Greenstadt (Brennan et. al, 2012) and Riddell-Juola (Wang et al., 2021) corpora.
- ❑ *Human (3)*: 870 news articles from 2021.
Sampled from PAN’24 training dataset.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset Creation (Training)

Dataset: “PAN AI Obfuscated 2025” (Training)

- ❑ *Human (1)*: 7,300 19th-century novels (500–700 words).
Scraped from Project Gutenberg.
- ❑ *Human (2)*: 931 essays.
Brennan-Greenstadt (Brennan et. al, 2012) and Riddell-Juola (Wang et al., 2021) corpora.
- ❑ *Human (3)*: 870 news articles from 2021.
Sampled from PAN’24 training dataset.
- ❑ *Machine*: Reconstruction of human texts by 14 LLMs (sampled 1.6:1).
Summarize and expand; GPT-3.5, GPT-4o, GPT-4o-mini, Gemini, DeepSeek, Llama, ...

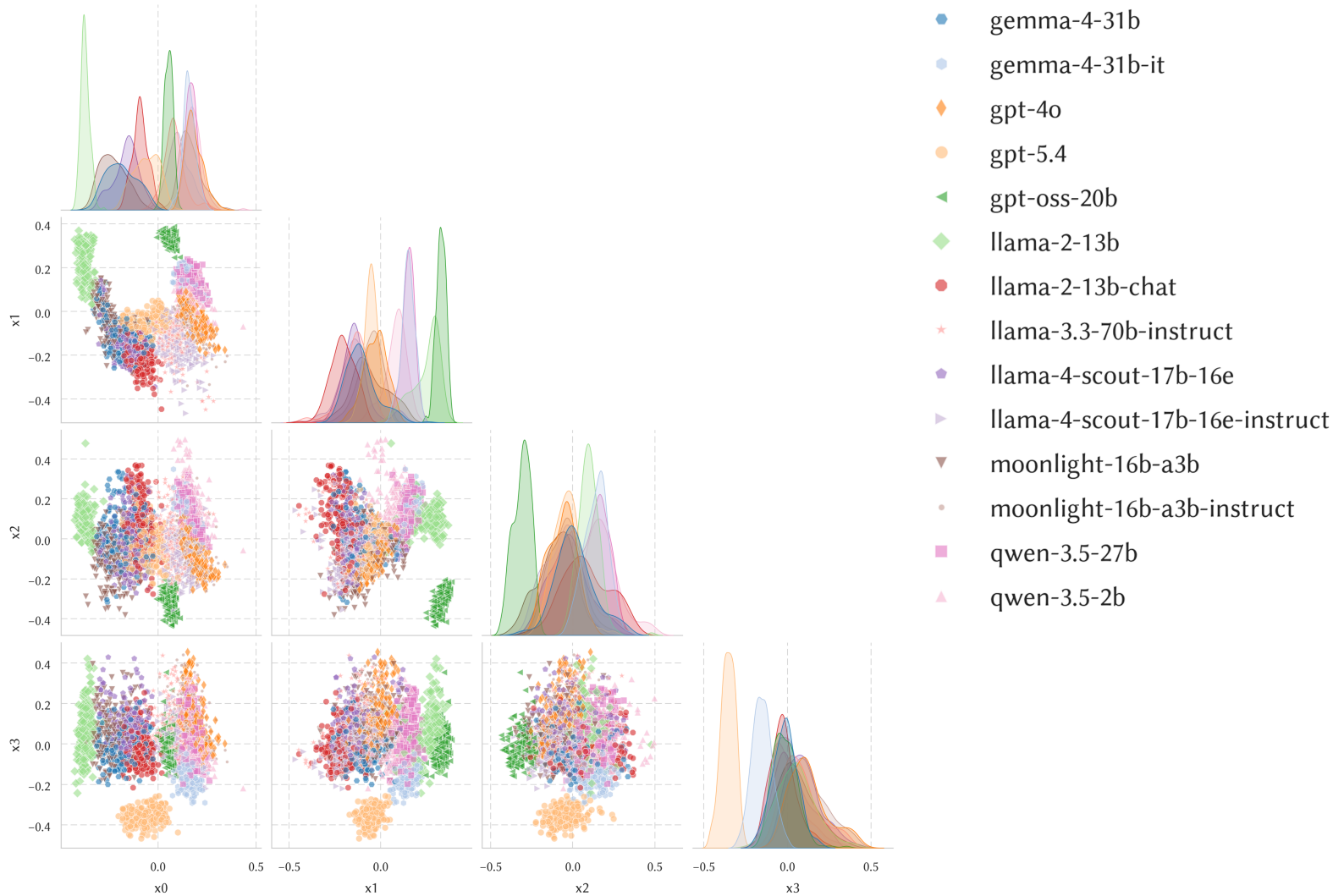
“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset Creation (Test)

What makes LLM text stand out?

- ❑ Post-training / alignment? [Xu et al., 2026]
- ❑ Just the right training data? [Bevendorff et al., 2025 / 2026]
- ❑ Both?

“Voigt-Kampff”^{*} Generative AI / LLM Detection



“Voight-Kampff”^{*} Generative AI / LLM Detection

Ways to suppress post-training style traits?

Conjecture: AI training and alignment datasets introduce "western" habits.

(See, e.g., the American “cheese” smile – [AI and the American Smile](#))

“Voight-Kampff”^{*} Generative AI / LLM Detection

Ways to suppress post-training style traits?

Conjecture: AI training and alignment datasets introduce "western" habits.

(See, e.g., the American “cheese” smile – [AI and the American Smile](#))

- ❑ Language encodes mental models and cultural practices.
- ❑ Social protocols expressed through language differ in non-western cultures.
- ❑ Non-latin scripts languages may foster different kinds of expression.
- ❑ Word order may influence mental models and processes.
- ❑ Low-resource languages challenge LLMs and NLP applications.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Ways to suppress post-training style traits?

Conjecture: AI training and alignment datasets introduce "western" habits.

(See, e.g., the American “cheese” smile – [AI and the American Smile](#))

- ❑ Language encodes mental models and cultural practices.
- ❑ Social protocols expressed through language differ in non-western cultures.
- ❑ Non-latin scripts languages may foster different kinds of expression.
- ❑ Word order may influence mental models and processes.
- ❑ Low-resource languages challenge LLMs and NLP applications.

↪ AI detectors fit on those traits. So, why not try something like... Tibetan?

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Source: New York Times – 200 Prompts for Argumentative Writing^{*}

- ❑ *Human:* 200 linked essays and articles.
- ❑ *Machine:* Responses to the following prompt:

Write an essay or op ed (about 500 words) about the topic "{{ topic }}" in Romani and then translate the result into Tibetan. Output only the final Tibetan version.

Translated back to English using Google Translate.

^{*}<https://archive.nytimes.com/learning.blogs.nytimes.com/2014/02/04/200-prompts-for-argumentative-writing/>

“Voight-Kampff” * Generative AI / LLM Detection

Initial Experiment I: GPTZero [https://app.gptzero.me/]

How Seriously Should We Take Standardized Tests?

Standardized tests should be taken seriously—but not too seriously. They can provide useful information about what students know, how schools are performing, and where extra support may be needed. But when too much importance is placed on a single test score, these exams can distort education rather than improve it. The main argument in favor of standardized tests is fairness. In theory, every student answers the same questions under the same conditions, making it easier to compare results across classrooms, schools, and districts. This can help identify achievement gaps and reveal whether certain groups of students are being underserved. Without some common measure, it becomes harder to know whether schools are truly preparing students with the same level of rigor.

Standardized tests can also offer accountability. Parents, policymakers, and the public understandably want to know whether schools are doing their job. Test results can highlight schools that need more resources or reforms. In some cases, testing data has exposed deep inequalities that might otherwise have been ignored. A low score can sometimes be a warning sign—not that a student has failed, but that the system around that student may be

AdvancedFeedbackPlagiarismHallucination

Basic ScanShareExport

AI

GPTZero AI DetectionModel 4.9b

We are highly confident this text was AI generated

Chance this entire text is...

AI 100%Mixed 0%Human 0%

Your most AI sentences

When schools, teachers, and students are...

See your most AI and human sentences

What's next?

Get a grade and feedback

Built by writing experts

Guidance for Teachers and reviewers

Text up-to-date

3,315 characters • 506 words

Re-scan

Proofread

“Voight-Kampff” * Generative AI / LLM Detection

Initial Experiment I: GPTZero [https://app.gptzero.me/]

How Seriously Should We Take Standardized Tests?

Standardized tests should be taken seriously—but not too seriously. They can provide useful information about what students know, how schools are performing, and where extra support may be needed. But when too much importance is placed on a single test score, these exams can distort education rather than improve it. The main argument in favor of standardized tests is fairness. In theory, every student answers the same questions under the same conditions, making it easier to compare results across classrooms, schools, and districts. This can help identify achievement gaps and reveal whether certain groups of students are being underserved. Without some common measure, it becomes harder to know whether schools are truly preparing students with the same level of rigor. Standardized tests can also offer accountability. Parents, policymakers, and the public understandably want to know whether schools are doing their job. Test results can highlight schools that need more resources or reforms. In some cases, testing data has exposed deep inequalities that might otherwise have been ignored. A low score can sometimes be a warning sign—not that a student has failed, but that the system around that student may be

Proofread

AdvancedFeedbackPlagiarismHallucination

Basic Scan

AI

We are highly confident this text was AI generated

Chance this entire text is... AI 100% Mixed 0% Human 0%

Your most AI sentences

When schools, teachers, and students are...

See your most AI and human sentences

Get a grade and feedback

Guidance for Teachers and reviewers

Text up-to-date

3,315 characters • 506 words

Re-scan

How Seriously Should We Take Standardized Tests?

Standardized tests should be taken seriously—but not too seriously. They can provide useful information about what students know, how schools are performing, and where extra support may be needed. But when too much importance is placed on a single test score, these exams can distort education rather than improve it. The main argument in favor of standardized tests is fairness. In theory, every student answers the same questions under the same conditions, making it easier to compare results across classrooms, schools, and districts. This can help identify achievement gaps and reveal whether certain groups of students are being underserved. Without some common measure, it becomes harder to know whether schools are truly preparing students with the same level of rigor. Standardized tests can also offer accountability. Parents, policymakers, and the public understandably want to know whether schools are doing their job. Test results can highlight schools that need more resources or reforms. In some cases, testing data has exposed deep inequalities that might otherwise have been ignored. A low score can sometimes be a warning sign—not that a student has failed, but that the system around that student may be

Proofread

AdvancedFeedbackPlagiarismHallucination

Basic Scan

human

We are highly confident this text is entirely human

Chance this entire text is... AI 4% Mixed 0% Human 96%

Your most AI sentences

When schools, teachers, and students are...

See your most AI and human sentences

Get a grade and feedback

Guidance for Teachers and reviewers

Text up-to-date

2,528 characters • 64 words

Re-scan

How Seriously Should We Take Standardized Tests?

Standardized tests should be taken seriously—but not too seriously. They can provide useful information about what students know, how schools are performing, and where extra support may be needed. But when too much importance is placed on a single test score, these exams can distort education rather than improve it. The main argument in favor of standardized tests is fairness. In theory, every student answers the same questions under the same conditions, making it easier to compare results across classrooms, schools, and districts. This can help identify achievement gaps and reveal whether certain groups of students are being underserved. Without some common measure, it becomes harder to know whether schools are truly preparing students with the same level of rigor. Standardized tests can also offer accountability. Parents, policymakers, and the public understandably want to know whether schools are doing their job. Test results can highlight schools that need more resources or reforms. In some cases, testing data has exposed deep inequalities that might otherwise have been ignored. A low score can sometimes be a warning sign—not that a student has failed, but that the system around that student may be

Proofread

AdvancedFeedbackPlagiarismHallucination

Basic Scan

human

We are highly confident this text is entirely human

Chance this entire text is... AI 4% Mixed 0% Human 96%

Your most AI sentences

When schools, teachers, and students are...

See your most AI and human sentences

Get a grade and feedback

Guidance for Teachers and reviewers

Text up-to-date

2,528 characters • 64 words

Re-scan

“Voight-Kampff”^{*} Generative AI / LLM Detection

Initial Experiment I: GPTZero [https://app.gptzero.me/]

How Seriously Should We Take Standardized Tests?

Standardized tests should be taken seriously—but not too seriously. They can provide useful information about what students know, how schools are performing, and where extra support may be needed. But when too much importance is placed on a single test score, these exams can distort education rather than improve it. The main argument in favor of standardized tests is fairness. In theory, every student answers the same questions under the same conditions, making it easier to compare results across classrooms, schools, and districts. This can help identify achievement gaps and reveal whether certain groups of students are being underserved. Without some common measure, it becomes harder to know whether schools are truly preparing students with the same level of rigor. Standardized tests can also offer accountability. Parents, policymakers, and the public understandably want to know whether schools are doing their job. Test results can highlight schools that need more resources or reforms. In some cases, testing data has exposed deep inequalities that might otherwise have been ignored. A low score can sometimes be a warning sign—not that a student has failed, but that the system around that student may be

AdvancedFeedbackPlagiarismHallucination

Basic ScanShareExport

AI

GPTZero AI DetectionModel 4.9b

We are highly confident this text was AI generated

Chance this entire text is...

AI 100%Mixed 0%Human 0%

See your most AI and human sentences

What's next?

Get a grade and feedback

Guidance for Teachers and reviewers

Text up-to-date3,315 characters • 506 wordsRe-scan

How serious should we be about standardized tests?

In today's educational world, Standardized examinations are based on students' abilities and skills. School quality It is used as a key tool to measure teachers' teaching responsibilities. Good grades will help you get into college. It is also considered an important factor in the development of education policy. but". The real question is this: How serious should we be about standardized tests?

The exam is not useless. first'. That would give all students the same scale. Schools in different locations You can also compare your knowledge between students from different categories. second'. It can highlight gaps and weaknesses in the education system. If a dataset from many schools shows a problem with basic math or reading, it's a good idea to have a school. It is a clear warning that correction is needed. On top of that Regardless of personal opinion or appearance Assessment by numbers and equivalence contributes to the legitimacy and accountability of education. nevertheless". If you give too much importance to standardized tests, There will be many big problems. First of all, Exams cannot measure a person's intelligence. Creativity analysis Quality of cooperation The will to face difficulties is important in life. but Many one-page questions

AdvancedFeedbackPlagiarismHallucination

Basic ScanShareExport

human

GPTZero AI DetectionModel 4.9b

We are moderately confident this text is entirely human

Chance this entire text is...

AI 29%Mixed 0%Human 71%

See sentence highlighting

Get a grade and feedback

Guidance for Teachers and reviewers

Text up-to-date2,465 characters • 399 wordsRe-scan

“Voight-Kampff”^{*} Generative AI / LLM Detection

Initial Experiment II: Pangram [<https://www.pangram.com/>]

AI Highlight

How Seriously Should We Take Standardized Tests?

Standardized tests should be taken seriously—but not too seriously. They can provide useful information about what students know, how schools are performing, and where extra support may be needed. But when too much importance is placed on a single test score, these exams can distort education rather than improve it.

The main argument in favor of standardized tests is fairness. In theory, every student answers the same questions under the same conditions, making it easier to compare results across classrooms, schools, and districts. This can help identify achievement gaps and reveal whether certain groups of students are being underserved. Without some common measure, it becomes harder to know whether schools are truly preparing students with the same level of rigor.

Standardized tests can also offer accountability. Parents, policymakers, and the public understandably want to know whether schools are doing their job. Test results can highlight schools that need more resources or reforms. In some cases, testing data has exposed deep inequalities that might otherwise have been ignored. A low score can sometimes be a warning sign—not that a student has failed, but that the system around that student may be failing.

Still, the problems begin when test scores are treated as the ultimate measure of learning. Education is far

AI Generated

518 words scanned

Pangram 4.0

100 %

of this text is AI

Upgrade

Share Result

“Voight-Kampff” * Generative AI / LLM Detection

Initial Experiment II: Pangram [https://www.pangram.com/]

AI Highlight

How Seriously Should We Take Standardized Tests?

Standardized tests should be taken seriously—but not too seriously. They can provide useful information about what students know, how schools are performing, and where extra support may be needed. But when too much importance is placed on a single test score, these exams can distort education rather than improve it.

The main argument in favor of standardized tests is fairness. In theory, every student answers the same questions under the same conditions, making it easier to compare results across classrooms, schools, and districts. This can help identify achievement gaps and reveal whether certain groups of students are being underserved. Without some common measure, it becomes harder to know whether schools are truly preparing students with the same level of rigor.

Standardized tests can also offer accountability. Parents, policymakers, and the public understandably want to know whether schools are doing their job. Test results can highlight schools that need more resources or reforms. In some cases, testing data has exposed deep inequalities that might otherwise have been ignored. A low score can sometimes be a warning sign—not that a student has failed, but that the system around that student may be failing.

Still, the problems begin when test scores are treated as the ultimate measure of learning. Education is far

AI Generated

518 words scanned

Pangram 4.0

100%
of this text is AI

Upgrade

Share Result

AI Highlight

ཚད་ལྡན་རྒྱལ་ཁེན་ལ་དང་ཚོས་གཞི་ཅན་གཟབ་ནན་བྱ་དགོས་སམ།

དེ་སང་གི་སློབ་གསོའི་འཛིན་རྟེན་ནང་། ཚད་ལྡན་རྒྱལ་ཁེན་ནི་སློབ་ཕྱག་གི་རུས་པ་དང་། སློབ་གྲྭའི་སྤྱི་བསྟན་གྱི་བསྟན་འགྲན་ལས་འགན་བཅས་འཇལ་བའི་གཙོ་ཆས་ཤིག་ཏུ་བཀོལ་བ་ཡིན། རྒྱལ་ཁེན་ལས་ཕན་པོ་ཐོབ་ན་མཐོ་སློབ་ཏུ་འཇུག་སྐབས་ཡན་ཐོགས་ཆེ་བ་དང་། སྤྱི་ཚན་ཁག་གིས་ཀྱང་སློབ་གསོའི་སྤྱི་ཚུལ་བཟོ་བའི་ཆ་ཤྭ་ན་གལ་ཆེན་ཞིག་ཏུ་བརྩི་བ་ཡིན། འོན་ཀྱང་། དྲི་བ་དངོས་ནི་འདི་ཡིན། དང་ཚོས་ཚད་ལྡན་རྒྱལ་ཁེན་ལ་གཞི་ཅན་གཟབ་ནན་བྱ་དགོས་སམ།

རྒྱལ་ཁེན་ལ་ཡན་ཡོན་མེད་པ་མིན། དང་པོ་དེས་སློབ་ཕྱག་ཚང་མར་འཇལ་ཚད་གཅིག་པ་ཞིག་སློབ་ཐུབ། ས་གནས་མི་འདྲ་བའི་སློབ་གྲྭ་དང་། རིགས་ཁག་མི་འདྲ་བའི་སློབ་ཕྱག་བར་ནས་ཀྱང་ཤེས་ཚད་བསྒྲུར་བྱེད་པ་ཡིན། གཉིས་པ་དེས་སློབ་གསོའི་ལམ་ལུགས་ནང་གི་སློབ་ཆ་དང་ཞན་ཆ་མངོན་པར་བྱེད་ཐུབ། གལ་ཏེ་སློབ་གྲྭ་མང་པོའི་གྲངས་ཐོ་ཞིག་གིས་གཞི་རྩའི་རྩིས་རིག་གམ་སློབ་ཆ་ལ་དཀའ་ངལ་ཡོད་པ་བསྟན་ན། དེ་ནི་བཟོ་བཅོས་དགོས་པའི་ཉེན་བརྟུན་ལས་པོ་ཞིག་ཡིན། དེའི་ཕྱིར་། མི་སློབ་ཀྱི་ལྷ་ཚུལ་དང་ཕྱི་ཚུལ་ལ་མ་བརྟེན་པར། གྲངས་ཀ་དང་མཚུངས་ཚད་ཀྱིས་བརྟེན་དཔྱད་བྱེད་པས་སློབ་གསོའི་ཁྲིམས་མཐུན་དང་འགན་ཁུར་ལ་ཡན་ཐོགས་ཡོད།

ཡིན་ནའང་། ཚད་ལྡན་རྒྱལ་ཁེན་ལ་ཚད་ལས་བསྟན་བའི་གལ་ཆེན་ཐུན་ན། དཀའ་ངལ་ཆེན་པོ་མང་པོ་འབྱུང་ངེས་ཡིན། ཐོག་མར། རྒྱལ་ཁེན་གྱིས་མི་ཞིག་གི་ཤེས་རབ་ཚད་མ་འཇལ་མི་ཐུབ། གསར་གཏོད་ཀྱི་རུས་པ། དཔྱད་ཤེས། མཉམ་ལས་ཀྱི་སྤྱི་ཚུལ་དང་། དཀའ་ངལ་ལ་གཞི་ཅན་ཕྱིར་པའི་བརྩོན་སེམས་བཅས་ནི་མི་ཆེད་ནང་གལ་ཆེན་ཡིན་ཡང་། ཤོག་ངོས་གཅིག་གི་དྲི་བ་མང་པོས་དེ་ཙམ་ཚད་མངོན་ཐུབ་མི་སྲིད། གཉིས་པ་སློབ་གྲྭ་ཁ་ཤས་ཀྱིས་“རྒྱལ་ཁེན་ཆེད་དུ་སློབ་པ་” ཞེས་པའི་ལམ་ལུགས་ལ་གཙོ་བོ་བསྟན་པས། སློབ་ཚན་ཡངས་པོ་དང་གོ་རྟོགས་ཟབ་པོ་ཡར་འབྱུངས་བྱས་ཏེ། དྲི་བའི་ལན་འདེབས་ཆུལ་ཁོ་ན་ལ་འབད་པ་ཡིན། དེས་སློབ་གསོ་ཞེས་པ་ནི་དོན་དམ་གྱི་ཤེས་ཡོན་འཛོལ་བ་མ་ཡིན་པར། ཅང་གྲངས་འཛོལ་བའི་ལས་ཀ་ཞིག་ཏུ་བསྒྲུར་རྟེན་ཡོང་།

Human Written

938 words scanned

Pangram 4.0

100%
of this text is Human Written

Upgrade

Share Result

“Voight-Kampff”^{*} Generative AI / LLM Detection

Initial Experiment II: Pangram [\[https://www.pangram.com/\]](https://www.pangram.com/)

AI Highlight

How Seriously Should We Take Standardized Tests?

Standardized tests should be taken seriously—but not too seriously. They can provide useful information about what students know, how schools are performing, and where extra support may be needed. But when too much importance is placed on a single test score, these exams can distort education rather than improve it.

The main argument in favor of standardized tests is fairness. In theory, every student answers the same questions under the same conditions, making it easier to compare results across classrooms, schools, and districts. This can help identify achievement gaps and reveal whether certain groups of students are being underserved. Without some common measure, it becomes harder to know whether schools are truly preparing students with the same level of rigor.

Standardized tests can also offer accountability. Parents, policymakers, and the public understandably want to know whether schools are doing their job. Test results can highlight schools that need more resources or reforms. In some cases, testing data has exposed deep inequalities that might otherwise have been ignored. A low score can sometimes be a warning sign—not that a student has failed, but that the system around that student may be failing.

Still, the problems begin when test scores are treated as the ultimate measure of learning. Education is far

AI Generated

518 words scanned

Pangram 4.0

100%
of this text is AI

Upgrade

Share Result

AI Highlight

How serious should we be about standardized tests?

In today's educational world, Standardized examinations are based on students' abilities and skills. School quality It is used as a key tool to measure teachers' teaching responsibilities. Good grades will help you get into college. It is also considered an important factor in the development of education policy. but". The real question is this: How serious should we be about standardized tests?

The exam is not useless. first'. That would give all students the same scale. Schools in different locations You can also compare your knowledge between students from different categories. second'. It can highlight gaps and weaknesses in the education system. If a dataset from many schools shows a problem with basic math or reading, it's a good idea to have a school. It is a clear warning that correction is needed. On top of that Regardless of personal opinion or appearance Assessment by numbers and equivalence contributes to the legitimacy and accountability of education.

nevertheless". If you give too much importance to standardized tests, There will be many big problems.

First of all, Exams cannot measure a person's intelligence. Creativity analysis Quality of cooperation The will to face difficulties is important in life, but Many one-page questions may not fully reflect them. second'.

AI Generated

408 words scanned

Pangram 4.0

100%
of this text is AI

This text appears to have been paraphrased or rewritten

Upgrade

Share Result

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Dataset: “PAN AI Obfuscated 2026”

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Dataset: “PAN AI Obfuscated 2026”

- ❑ *Human*: 200 NYTimes articles and essays.
Scraped from New York Times Writing Prompts.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Dataset: “PAN AI Obfuscated 2026”

- ❑ *Human*: 200 NYTimes articles and essays.
Scraped from New York Times Writing Prompts.
- ❑ *Machine (1)*: Writing prompts Tibetan.
GPT-5.5, 5.4, 4o, Llama 3.3, translated via Google Translate.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Dataset: “PAN AI Obfuscated 2026”

- ❑ *Human*: 200 NYTimes articles and essays.
Scraped from New York Times Writing Prompts.
- ❑ *Machine (1)*: Writing prompts Tibetan.
GPT-5.5, 5.4, 4o, Llama 3.3, translated via Google Translate.
- ❑ *Machine (2)*: Writing prompts Tibetan round-trips.
Translated via Google Translate through multiple low-resource round-trips.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Dataset: “PAN AI Obfuscated 2026”

- ❑ *Human*: 200 NYTimes articles and essays.
Scraped from New York Times Writing Prompts.
- ❑ *Machine (1)*: Writing prompts Tibetan.
GPT-5.5, 5.4, 4o, Llama 3.3, translated via Google Translate.
- ❑ *Machine (2)*: Writing prompts Tibetan round-trips.
Translated via Google Translate through multiple low-resource round-trips.
- ❑ *Machine (3)*: Reverse Swahili.
GPT 5.5: Output in Swahili, but backwards. Then reversed in code.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Dataset: “PAN AI Obfuscated 2026”

- ❑ *Machine (4)*: Logit biasing.

Top-k blacklist sampling. Extreme KGW watermarking [Kirchenbauer et al, 2023].

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Dataset: “PAN AI Obfuscated 2026”

- ❑ *Machine (4)*: Logit biasing.
Top-k blacklist sampling. Extreme KGW watermarking [Kirchenbauer et al, 2023].
- ❑ *Machine (5)*: Heuristic authorship obfuscation.
Heuristic text obfuscation [Bevendorff et al., 2019].

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Dataset: “PAN AI Obfuscated 2026”

- ❑ *Machine (4)*: Logit biasing.
Top-k blacklist sampling. Extreme KGW watermarking [Kirchenbauer et al, 2023].
- ❑ *Machine (5)*: Heuristic authorship obfuscation.
Heuristic text obfuscation [Bevendorff et al., 2019].
- ❑ *Human (2)*: Brown corpus sample.
Out-of-domain control set.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Dataset: “PAN AI Obfuscated 2026”

- ❑ *Machine (4)*: Logit biasing.
Top-k blacklist sampling. Extreme KGW watermarking [Kirchenbauer et al, 2023].
- ❑ *Machine (5)*: Heuristic authorship obfuscation.
Heuristic text obfuscation [Bevendorff et al., 2019].
- ❑ *Human (2)*: Brown corpus sample.
Out-of-domain control set.
- ❑ *Machine (6)*: Brown corpus continuations.
Out-of-domain control set prompted on Brown corpus prefixes.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset (Test)

Dataset: “PAN AI Obfuscated 2026”

- ❑ *Machine (4)*: Logit biasing.
Top-k blacklist sampling. Extreme KGW watermarking [Kirchenbauer et al, 2023].
- ❑ *Machine (5)*: Heuristic authorship obfuscation.
Heuristic text obfuscation [Bevendorff et al., 2019].
- ❑ *Human (2)*: Brown corpus sample.
Out-of-domain control set.
- ❑ *Machine (6)*: Brown corpus continuations.
Out-of-domain control set prompted on Brown corpus prefixes.

Sampling: All texts truncated to about 4,000 characters and sampled to a 1:3 class ratio of *human* and *machine*.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Baselines

We provided 5 baseline systems:

- ❑ Mdok – 2025 winner [Macko, 2025]
- ❑ Binoculars [Hans et al., 2024]
- ❑ PPMd Compression-based Cosine [Sculley and Brodly, 2006; Halvani et al., 2017]
- ❑ SVM with TF-IDF features
- ❑ Zero-knowledge (always predict majority class)

“Voight-Kampff”^{*} Generative AI / LLM Detection

Submission Evaluation

- ❑ Submissions

34 teams (27 notebooks)

- ❑ Evaluation Measures

ROC-AUC, Brier, C@1, $F_{0.5u}$, F_1 , Mean of all

“Voight-Kampff”^{*} Generative AI / LLM Detection

Submission Evaluation

- ❑ Submissions

34 teams (27 notebooks)

- ❑ Evaluation Measures

ROC-AUC, Brier, C@1, $F_{0.5u}$, F_1 , Mean of all

- ❑ Winning System

Weighted ensemble of two fine-tuned Qwen models.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Submission Evaluation

- ❑ Submissions
34 teams (27 notebooks)
- ❑ Evaluation Measures
ROC-AUC, Brier, C@1, $F_{0.5u}$, F_1 , Mean of all
- ❑ Winning System
Weighted ensemble of two fine-tuned Qwen models.
- ❑ Runners-up
Bayesian BERT model with contrastive loss.

“Voight-Kampff”^{*} Generative AI / LLM Detection

Submission Evaluation

- ❑ Submissions

34 teams (27 notebooks)

- ❑ Evaluation Measures

ROC-AUC, Brier, C@1, $F_{0.5u}$, F_1 , Mean of all

- ❑ Winning System

Weighted ensemble of two fine-tuned Qwen models.

- ❑ Runners-up

Bayesian BERT model with contrastive loss.

- ❑ Other popular approaches

Fine-tuned *BERT, stylometric features, data augmentation.

“Voight-Kampff”^{*} Generative AI / LLM Detection

System Approaches (Overview 2025)

- ❑ LLM / PLM Embeddings

13 systems

- ❑ Text Perplexity

3 systems

- ❑ Term Frequencies

3 systems

- ❑ Stylometric Features

7 systems

- ❑ Ensemble Methods

5 systems

- ❑ Augmented Data

6 systems

- ❑ Custom Loss

5 systems

“Voight-Kampff”^{*} Generative AI / LLM Detection

System Approaches (Overview 2026)

- ❑ LLM / PLM Embeddings

21 systems

- ❑ Text Perplexity

4 systems

- ❑ Term Frequencies

10 systems

- ❑ Stylometric Features

10 systems

- ❑ Ensemble Methods

8 systems

- ❑ Augmented Data

10 systems

- ❑ Custom Loss

4 systems

“Voight-Kampff”^{*} Generative AI / LLM Detection

Systems Ranking (Top 13)

Team	ROC-AUC	Brier	C@1	F ₁	F _{0.5u}	Mean↓	FNR	FPR
1 suspiciously-coherent	0.989	0.965	0.965	0.976	0.980	0.975	0.053	0.029
2 dactyl	0.993	0.970	0.962	0.975	0.969	0.974	0.108	0.014
3 plagiarism-detector	0.997	0.958	0.949	0.955	0.982	0.968	0.000	0.067
<i>Baseline: mdok, winner 2025</i>	0.995	0.951	0.946	0.963	0.984	0.968	0.002	0.072
4 dsgt-pan-26	0.981	0.885	0.873	0.907	0.961	0.921	0.000	0.170
5 ydong	0.930	0.915	0.886	0.928	0.896	0.911	0.420	0.014
6 turingteam	0.957	0.882	0.846	0.851	0.925	0.892	0.053	0.179
7 writerslogic	0.891	0.891	0.853	0.902	0.899	0.887	0.312	0.092
8 a-a-detection	0.852	0.848	0.826	0.885	0.880	0.858	0.375	0.106
<i>Baseline: PPMd</i>	0.783	0.810	0.783	0.865	0.828	0.814	0.728	0.066
9 ak-ys	0.911	0.754	0.723	0.774	0.891	0.811	0.016	0.365
10 ikr3	0.814	0.779	0.749	0.856	0.788	0.797	1.000	0.000
11 radar	0.772	0.805	0.722	0.800	0.839	0.788	0.339	0.258
12 sinai	0.624	0.771	0.748	0.856	0.788	0.757	1.000	0.001
<i>Baseline: Binoculars (Llama-3.1)</i>	0.764	0.778	0.634	0.692	0.823	0.738	0.125	0.414
13 egrantha-ai	0.500	0.749	0.749	0.856	0.788	0.728	1.000	0.000
<i>Baseline: Zero-knowledge</i>	0.500	0.749	0.749	0.856	0.788	0.728	1.000	0.000
⋮								

“Voight-Kampff”^{*} Generative AI / LLM Detection

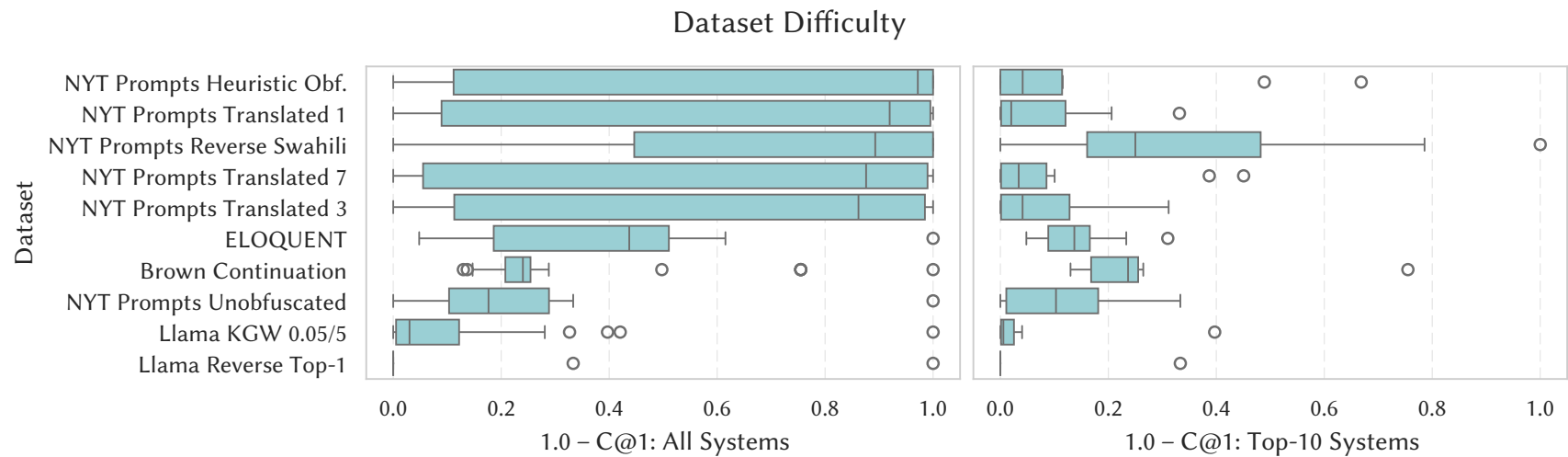
Systems Ranking (Top 13)

Team	ROC-AUC	Brier	C@1	F ₁	F _{0.5u}	Mean↓	FNR	FPR
1 suspiciously-coherent	0.989	0.965	0.965	0.976	0.980	0.975	0.053	0.029
2 dactyl	0.993	0.970	0.962	0.975	0.969	0.974	0.108	0.014
3 plagiarism-detector	0.997	0.958	0.949	0.955	0.982	0.968	0.000	0.067
Baseline: mdok, winner 2025	0.995	0.951	0.946	0.963	0.984	0.968	0.002	0.072
4 dsgrt-pan-26	0.981	0.885	0.873	0.907	0.961	0.921	0.000	0.170
5 ydong	0.930	0.915	0.886	0.928	0.896	0.911	0.420	0.014
6 turingteam	0.957	0.882	0.846	0.851	0.925	0.892	0.053	0.179
7 writerslogic	0.891	0.891	0.853	0.902	0.899	0.887	0.312	0.092
8 a-a-detection	0.852	0.848	0.826	0.885	0.880	0.858	0.375	0.106
Baseline: PPMd	0.783	0.810	0.783	0.865	0.828	0.814	0.728	0.066
9 ak-ys	0.911	0.754	0.723	0.774	0.891	0.811	0.016	0.365
10 ikr3	0.814	0.779	0.749	0.856	0.788	0.797	1.000	0.000
11 radar	0.772	0.805	0.722	0.800	0.839	0.788	0.339	0.258
12 sinai	0.624	0.771	0.748	0.856	0.788	0.757	1.000	0.001
Baseline: Binoculars (Llama-3.1)	0.764	0.778	0.634	0.692	0.823	0.738	0.125	0.414
13 egrantha-ai	0.500	0.749	0.749	0.856	0.788	0.728	1.000	0.000
Baseline: Zero-knowledge	0.500	0.749	0.749	0.856	0.788	0.728	1.000	0.000
⋮								

“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset Difficulty (Inverse C@1)

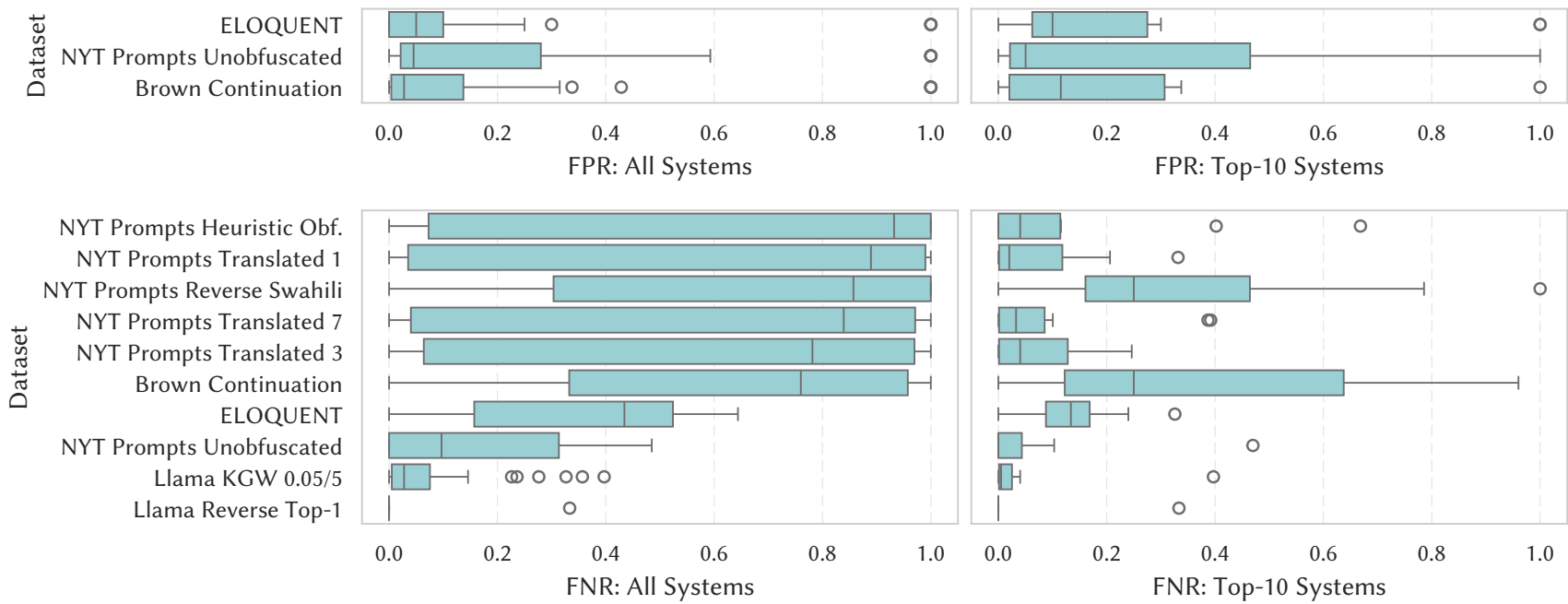
Dataset Difficulty as inverse C@1 over data subsets (all vs. top-10 systems):



“Voight-Kampff”^{*} Generative AI / LLM Detection

Dataset Difficulty (FPR / FNR)

False positive / negative rates (FPR / FNR) over data subsets (all vs. top-10 systems):
(positive = machine)



PAN 2026

Lab on Digital Text Forensics and Stylometry

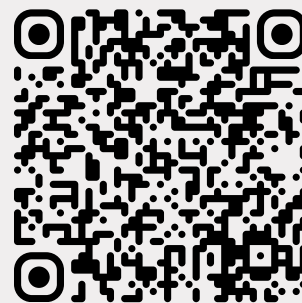


pan.webis.de

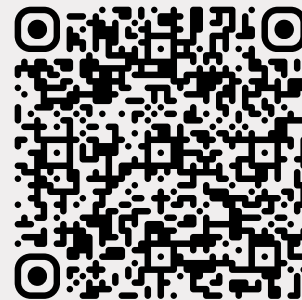


pan@webis.de

Thanks!



Task Website



GitHub Repository