

PAN 2024

Lab on Digital Text Forensics and Stylometry



pan.webis.de



pan@webis.de

Overview: “Voight-Kampff” Generative AI Authorship Verification

Janek Bevendorff

Matti Wiegmann

Jussi Karlgren

Luise Dürlich

Evangelia Gogoulou

Aarne Talman

Efstathios Stamatatos

Martin Potthast

Benno Stein

Leipzig University

Bauhaus-Universität Weimar

Silo AI

RISE Research Institutes of Sweden

University of Helsinki

University of the Aegean

University of Kassel

hessian.AI

ScaDS.AI

Voight-Kampff* Generative AI Authorship Verification

*Given two texts, one written by a human, the other by a large language model:
decide which text was written by whom.*

* 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 Authorship Verification

*Given two texts, one written by a human, the other by a large language model:
decide which text was written by whom.*

Organized as builder-breaker evaluation between two labs:

- ❑ **Builder** (30 Submissions)

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

- ❑ **Breaker** (4 Submissions)

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 Authorship Verification

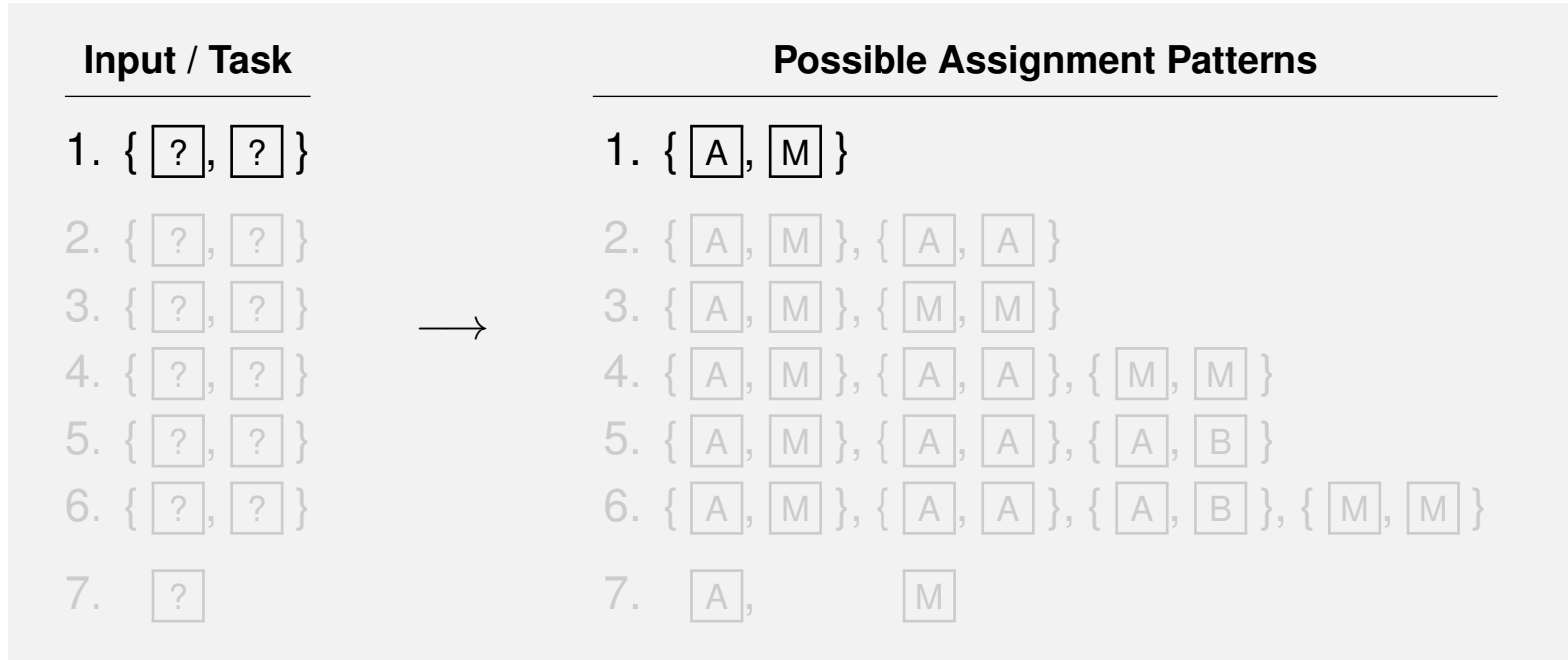
Task Formulation

Input / Task	Possible Assignment Patterns
1. { [?], [?] }	1. { [A], [M] }
2. { [?], [?] }	2. { [A], [M] }, { [A], [A] }
3. { [?], [?] }	3. { [A], [M] }, { [M], [M] }
4. { [?], [?] }	4. { [A], [M] }, { [A], [A] }, { [M], [M] }
5. { [?], [?] }	5. { [A], [M] }, { [A], [A] }, { [A], [B] }
6. { [?], [?] }	6. { [A], [M] }, { [A], [A] }, { [A], [B] }, { [M], [M] }
7. [?]	7. [A], [M]

([A] / [B] : Human authors, [M] : Machine)

Voight-Kampff Generative AI Authorship Verification

Task Formulation



([A] / [B]: Human authors, [M]: Machine)

Voight-Kampff Generative AI Authorship Verification

Dataset Creation

Dataset: “PAN AI News 2021”

- ❑ *Human text:* 1,359 US news article from 2021.
Crawled From Google News
- ❑ *Machine text:* Reconstruction of source texts by 9 (13) LLMs.
Summarize and expand; GPT, Gemini, Llama, Alpaca, . . .

Voight-Kampff Generative AI Authorship Verification

Dataset Creation (Prompt)

Summary You are a news article and press release summarizer. Given an article, you summarize the key points in 10 bullet points.

Voight-Kampff Generative AI Authorship Verification

Dataset Creation (Prompt)

Summary You are a news article and press release summarizer. Given an article, you summarize the key points in 10 bullet points.

Type You also classify the article type ("breaking news", "press release", "government agency statement", "financial news", "opinion piece", "fact check", "celebrity news", "general reporting", "speech transcript").

Voight-Kampff Generative AI Authorship Verification

Dataset Creation (Prompt)

- Summary** You are a news article and press release summarizer. Given an article, you summarize the key points in 10 bullet points.
- Type** You also classify the article type ("breaking news", "press release", "government agency statement", "financial news", "opinion piece", "fact check", "celebrity news", "general reporting", "speech transcript").
- Dateline** Extract the dateline from the beginning of the article if one exists (e.g. "Washington" or "May 28 (Reuters)").

Voigt-Kampff Generative AI Authorship Verification

Dataset Creation (Prompt)

- Summary** You are a news article and press release summarizer. Given an article, you summarize the key points in 10 bullet points.
- Type** You also classify the article type ("breaking news", "press release", "government agency statement", "financial news", "opinion piece", "fact check", "celebrity news", "general reporting", "speech transcript").
- Dateline** Extract the dateline from the beginning of the article if one exists (e.g. "Washington" or "May 28 (Reuters)").
- Quotes** If spokespersons are cited verbatim, list their names, functions, and titles (if any).

Voight-Kampff Generative AI Authorship Verification

Dataset Creation (Prompt)

- Summary** You are a news article and press release summarizer. Given an article, you summarize the key points in 10 bullet points.
- Type** You also classify the article type ("breaking news", "press release", "government agency statement", "financial news", "opinion piece", "fact check", "celebrity news", "general reporting", "speech transcript").
- Dateline** Extract the dateline from the beginning of the article if one exists (e.g. "Washington" or "May 28 (Reuters)").
- Quotes** If spokespersons are cited verbatim, list their names, functions, and titles (if any).
- Audience** Determine the article's target audience ("general public", "professionals", "children").

Voight-Kampff Generative AI Authorship Verification

Dataset Creation (Prompt)

- Summary** You are a news article and press release summarizer. Given an article, you summarize the key points in 10 bullet points.
- Type** You also classify the article type ("breaking news", "press release", "government agency statement", "financial news", "opinion piece", "fact check", "celebrity news", "general reporting", "speech transcript").
- Dateline** Extract the dateline from the beginning of the article if one exists (e.g. "Washington" or "May 28 (Reuters)").
- Quotes** If spokespersons are cited verbatim, list their names, functions, and titles (if any).
- Audience** Determine the article's target audience ("general public", "professionals", "children").
- Stance** Classify whether the article's stance is "left-leaning", "right-leaning", or "neutral".

Voight-Kampff Generative AI Authorship Verification

Dataset Creation (Prompt)

Summary You are a news article and press release summarizer. Given an article, you summarize the key points in 10 bullet points.

Type You also classify the article type ("breaking news", "press release", "government agency statement", "financial news", "opinion piece", "fact check", "celebrity news", "general reporting", "speech transcript").

Dateline Extract the dateline from the beginning of the article if one exists (e.g. "Washington" or "May 28 (Reuters)").

Quotes If spokespersons are cited verbatim, list their names, functions, and titles (if any).

Audience Determine the article's target audience ("general public", "professionals", "children").

Stance Classify whether the article's stance is "left-leaning", "right-leaning", or "neutral".

Structure Answer in structured JSON format (without Markdown formatting) like so:

```
{
  "key_points": ["key point 1", "key point 2", ...],
  "spokespersons": ["person1 (title, function)", ...],
  "article_type": "article type",
  "dateline": "dateline",
  "audience": "audience",
  "stance": "stance"
}
```

Voight-Kampff Generative AI Authorship Verification

Dataset Creation (Prompting and Cleaning)

1. All LLMs were prompted with a template based on the summaries.

```
You are a {{ publisher }} journalist writing {{ article_type }}.
```

```
In your article, cover the following key points: ...
```

Voight-Kampff Generative AI Authorship Verification

Dataset Creation (Prompting and Cleaning)

1. All LLMs were prompted with a template based on the summaries.

```
You are a {{ publisher }} journalist writing {{ article_type }}.  
In your article, cover the following key points: ...
```

2. The generated texts were cleaned manually of artifacts, such as:

- “Sure, I’d be happy to help.”*
- “Sorry, I cannot. . . ”*
- “Here’s your article:”*
- “Here are 10 paragraphs:”*
- “In this article, I will. . . ”*
- placeholders such as *“[your name]”, “[email]”, “[end of article]”*
- paragraph numbers, bullet points, word “counts”
- etc.

Voight-Kampff Generative AI Authorship Verification

Dataset Creation

- ❑ *Human text*: 1,359 US news article from 2021.
Crawled From Google News
- ❑ *Machine text*: Reconstruction of source texts by 9 (13) LLMs.
Summarize and expand; GPT, Gemini, Llama, Alpaca, ...

Voight-Kampff Generative AI Authorship Verification

Dataset Creation

- ❑ *Human text*: 1,359 US news article from 2021.
Crawled From Google News
- ❑ *Machine text*: Reconstruction of source texts by 9 (13) LLMs.
Summarize and expand; GPT, Gemini, Llama, Alpaca, ...
- ❑ *Test data*: 3,411 pairs of human and AI text.

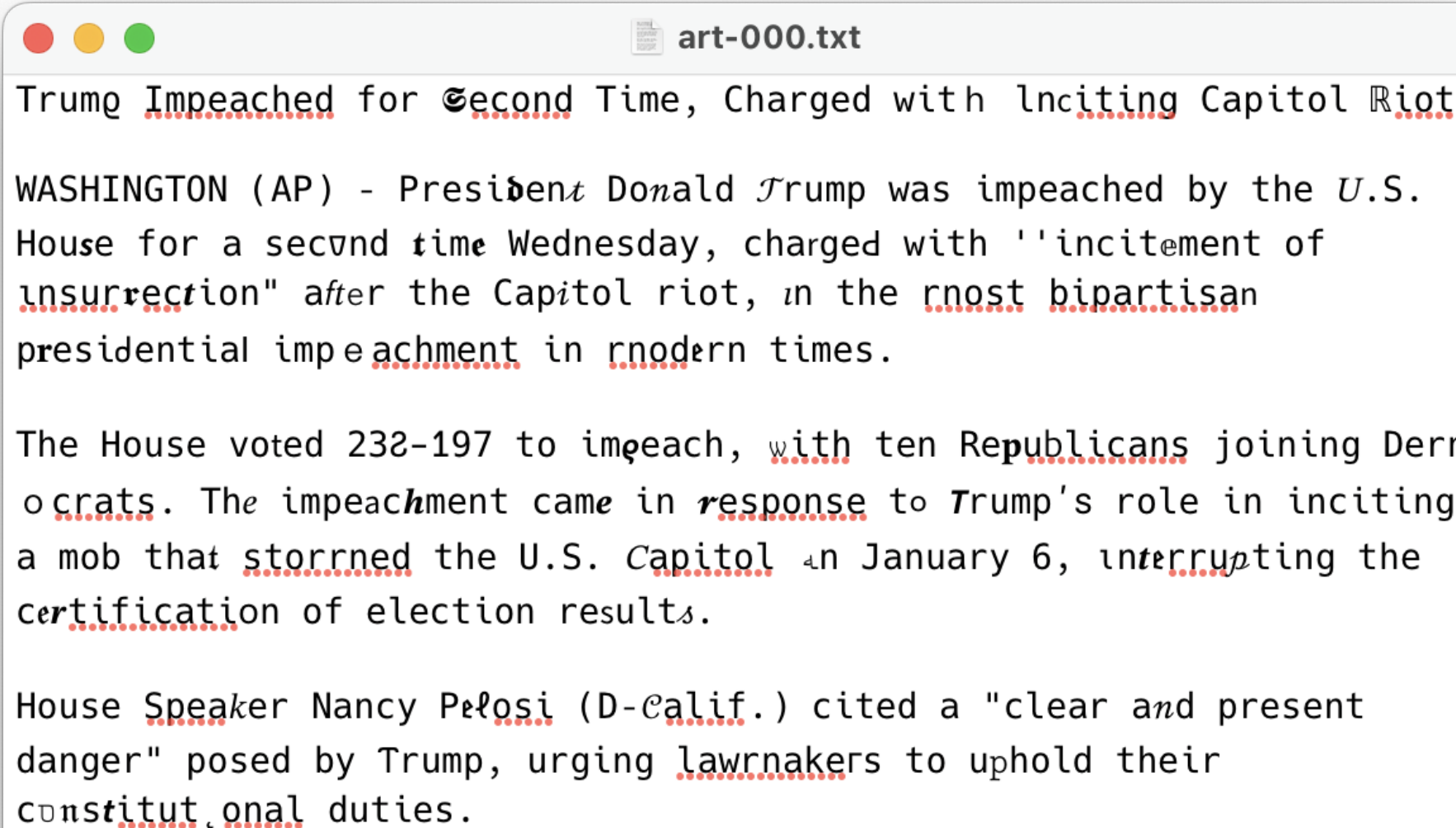
Voight-Kampff Generative AI Authorship Verification

Dataset Creation

- ❑ *Human text*: 1,359 US news article from 2021.
Crawled From Google News
- ❑ *Machine text*: Reconstruction of source texts by 9 (13) LLMs.
Summarize and expand; GPT, Gemini, Llama, Alpaca, ...
- ❑ *Test data*: 3,411 pairs of human and AI text.
- + 6 (7) dataset variations.
Unicode obfuscation, cross-topic pairs, contrastive decoding, cropped text (35 words), paraphrase prompt, cross-language pairs.
- + 5 ELOQUENT datasets (incl. baseline).

Voight-Kampff Generative AI Authorship Verification

Unicode Obfuscation (Example)



art-000.txt

Trump Impeached for Second Time, Charged with Inciting Capitol Riot

WASHINGTON (AP) - President Donald Trump was impeached by the U.S. House for a second time Wednesday, charged with "incitement of insurrection" after the Capitol riot, in the most bipartisan presidential impeachment in modern times.

The House voted 232-197 to impeach, with ten Republicans joining Democrats. The impeachment came in response to Trump's role in inciting a mob that stormed the U.S. Capitol in January 6, interrupting the certification of election results.

House Speaker Nancy Pelosi (D-Calif.) cited a "clear and present danger" posed by Trump, urging lawmakers to uphold their constitutional duties.

Voight-Kampff Generative AI Authorship Verification

Baselines

We provided 13 baseline systems:

- ❑ 2× each with Falcon-7B and Mistral-7B:
 - DetectGPT [Mitchell et al., 2023]
 - Fast-DetectGPT [Bao et al., 2023]
 - DetectLLM LRR and NPR [Su et al., 2023]
 - Binoculars [Hans et al., 2024]
- ❑ PPMd Compression-based Cosine [Sculley and Brodly, 2006; Halvani et al., 2017]
- ❑ Authorship Unmasking [Koppel and Schler, 2004; Bevendorff et al., 2019]
- ❑ Text length comparison

Voight-Kampff Generative AI Authorship Verification

Submission Evaluation

- ❑ Submissions
30 total (27 notebooks)
- ❑ Evaluation Measures
ROC-AUC, Brier, C@1, $F_{0.5u}$, F_1 , Mean of all

Voight-Kampff Generative AI Authorship Verification

Submission Evaluation

- ❑ Submissions
30 total (27 notebooks)
- ❑ Evaluation Measures
ROC-AUC, Brier, C@1, $F_{0.5u}$, F_1 , Mean of all
- ❑ Winning System
Ensemble of Binoculars and a fine-tuned Mistral + Llama

Voight-Kampff Generative AI Authorship Verification

Submission Evaluation

- ❑ Submissions
30 total (27 notebooks)
- ❑ Evaluation Measures
ROC-AUC, Brier, C@1, $F_{0.5u}$, F_1 , Mean of all
- ❑ Winning System
Ensemble of Binoculars and a fine-tuned Mistral + Llama
- ❑ Runners-up
“Tri-sentence analysis” with MPU loss,* SVM with TF-IDF features

* <https://arxiv.org/abs/2305.18149>

Voight-Kampff Generative AI Authorship Verification

Submission Evaluation

- ❑ Submissions
30 total (27 notebooks)
- ❑ Evaluation Measures
ROC-AUC, Brier, C@1, $F_{0.5u}$, F_1 , Mean of all
- ❑ Winning System
Ensemble of Binoculars and a fine-tuned Mistral + Llama
- ❑ Runners-up
“Tri-sentence analysis” with MPU loss,^{*} SVM with TF-IDF features
- ❑ Other popular approaches
Fine-tuned BERT, perplexity, (various) ensembles.

^{*}<https://arxiv.org/abs/2305.18149>

Voight-Kampff Generative AI Authorship Verification

System Approaches (Overview)

- ❑ LLM / PLM Embeddings

20 systems

- ❑ Text Perplexity

11 systems

- ❑ Term Frequencies

1 system

- ❑ Stylometric Features

5 systems

- ❑ Ensemble Methods

5 systems

- ❑ Augmented Data

6 systems

- ❑ Zero-shot

0 systems

Voight-Kampff Generative AI Authorship Verification

Systems Ranking (Main Test Dataset Only)

	Team	ROC-AUC	Brier	C@1	F ₁	F _{0.5u}	Mean
1	Tavan	0.999	0.990	0.993	0.993	0.997	0.995
2	Valdez-Valenzuela	0.985	0.985	0.985	0.985	0.983	0.985
3	Zi. Lin	0.979	0.979	0.979	0.979	0.980	0.979
4	J. Huang	0.980	0.980	0.980	0.979	0.977	0.979
5	L. Guo	0.979	0.963	0.947	0.947	0.945	0.957
			⋮				
9	Lorenz	0.973	0.898	0.952	0.951	0.950	0.946
	<i>Binoculars (Falcon 7B)</i>	0.943	0.928	0.926	0.920	0.922	0.928
			⋮				
	<i>DetectGPT (Falcon 7B)</i>	0.493	0.663	0.489	0.487	0.487	0.525
			⋮				

(Scores discounted by 0.5σ)

Voight-Kampff Generative AI Authorship Verification

Systems Ranking (Final)

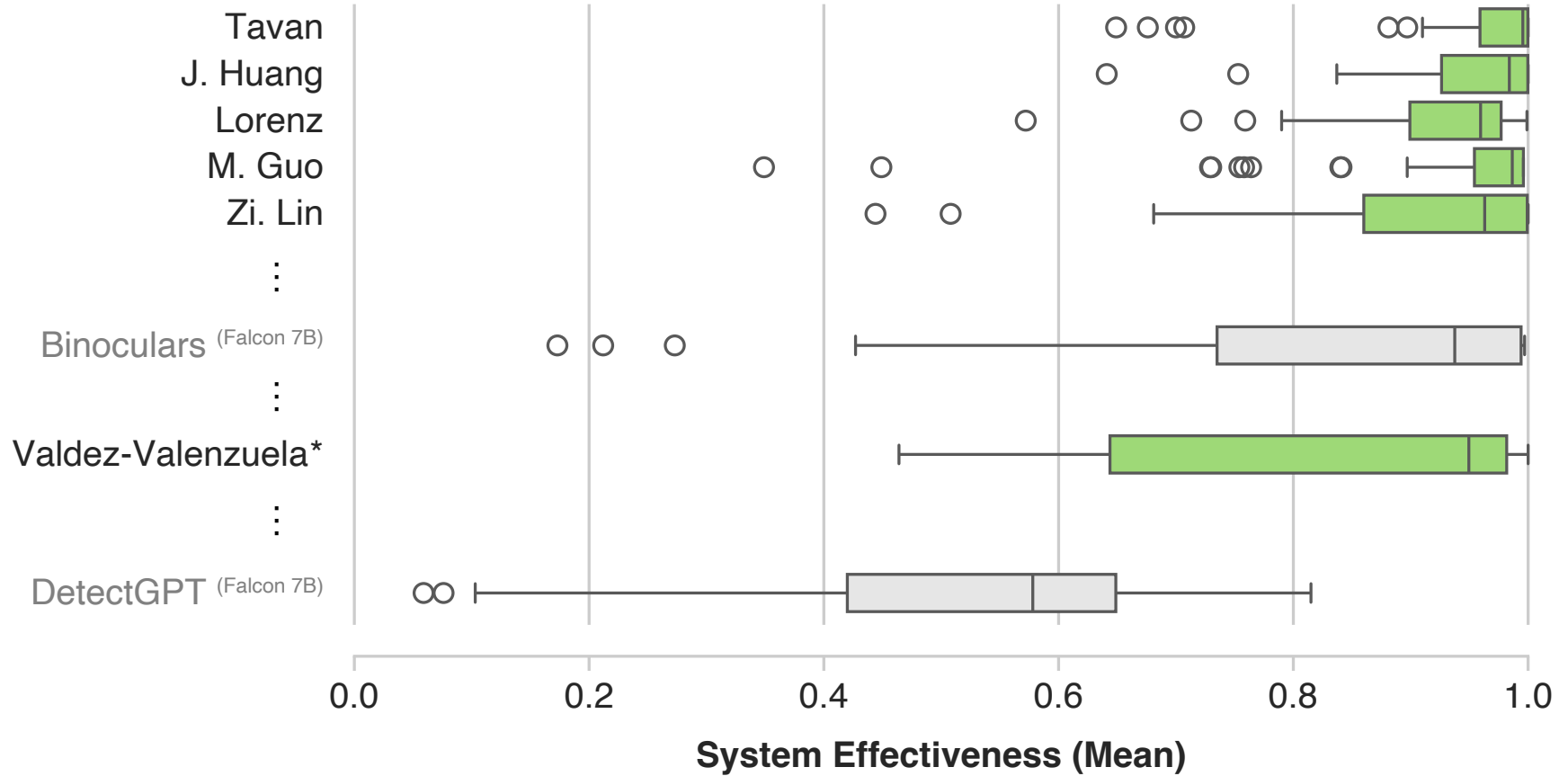
	Team	ROC-AUC	Brier	C@1	F ₁	F _{0.5u}	Mean
1	Tavan	0.961	0.928	0.912	0.884	0.932	0.924
2	J. Huang	0.931	0.926	0.928	0.905	0.913	0.921
3	Lorenz	0.925	0.869	0.882	0.875	0.869	0.886
4	M. Guo	0.889	0.875	0.887	0.884	0.884	0.884
5	Zi. Lin	0.851	0.850	0.850	0.852	0.849	0.851
			⋮				
	<i>Binoculars (Falcon 7B)</i>	0.751	0.780	0.734	0.720	0.720	0.741
			⋮				
14	Valdez-Valenzuela	0.741 *	0.760 *	0.718 *	0.711 *	0.695 *	0.727 *
			⋮				
	<i>DetectGPT (Falcon 7B)</i>	0.409	0.526	0.425	0.413	0.412	0.439

(Scores discounted by 0.5σ)

* Scores estimated due to run failures on short texts.

Voight-Kampff Generative AI Authorship Verification

Submission Score Distribution (Final)



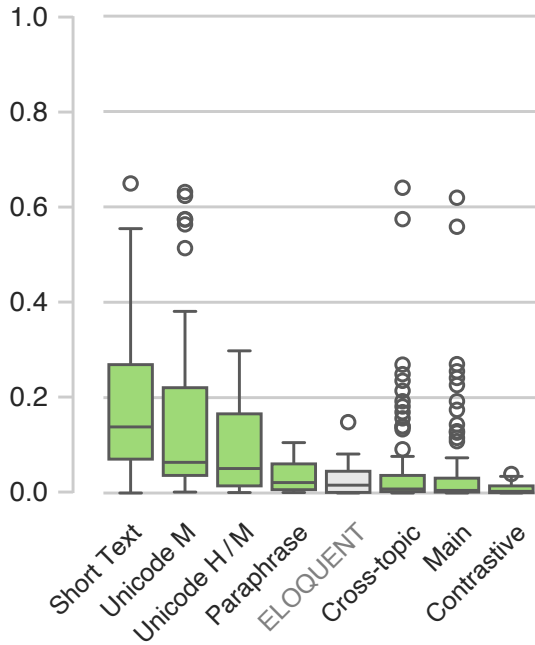
* Scores estimated due to run failures on short texts.

Voight-Kampff Generative AI Authorship Verification

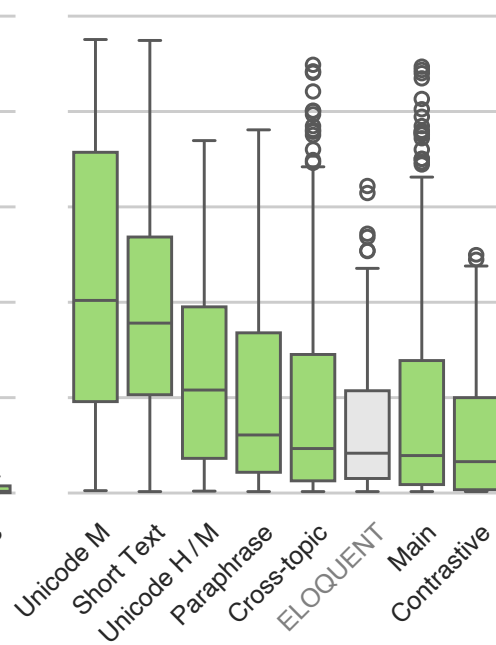
Dataset Difficulty

Dataset Difficulty PAN

10 Best Systems



All Systems

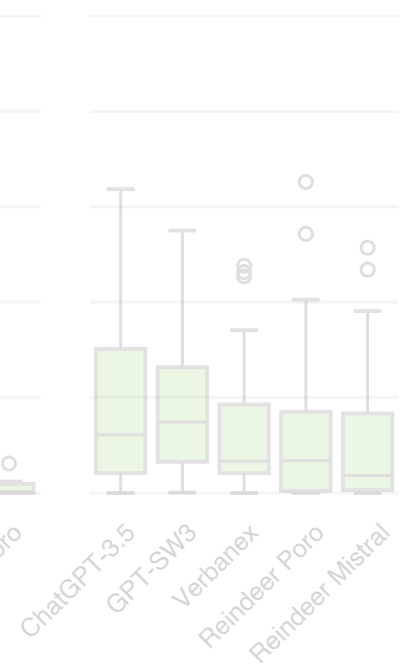


Dataset Difficulty ELOQUENT

10 Best Systems



All Systems

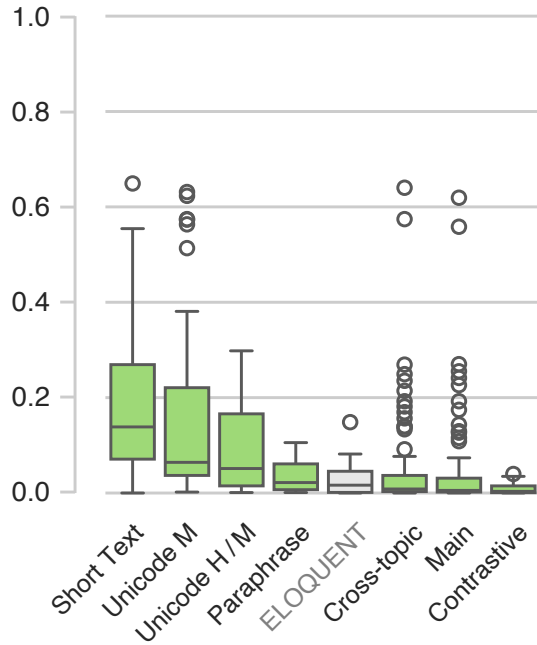


Voight-Kampff Generative AI Authorship Verification

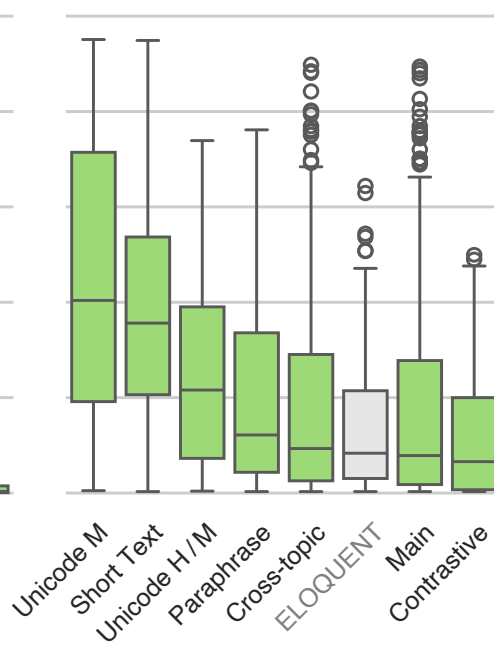
Dataset Difficulty

Dataset Difficulty PAN

10 Best Systems

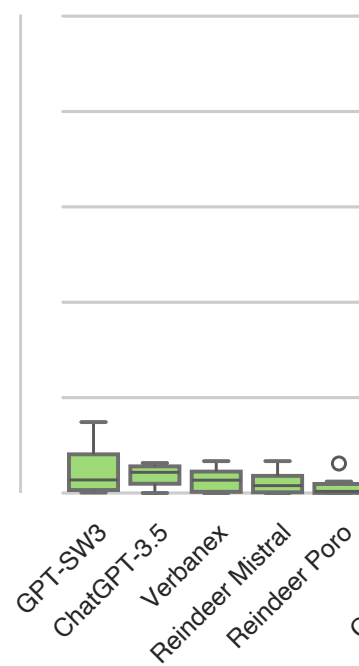


All Systems

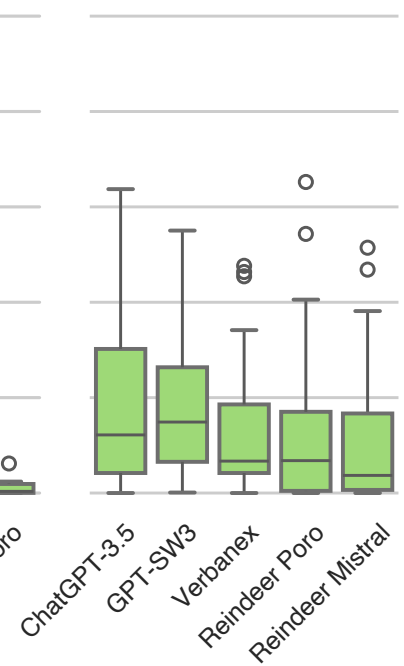


Dataset Difficulty ELOQUENT

10 Best Systems



All Systems

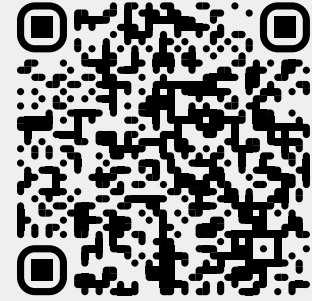


PAN 2024

Lab on Digital Text Forensics and Stylometry

 pan.webis.de  pan@webis.de

Thanks!



[Task Website](#)



[GitHub Repository](#)