JAGIELLONIAN UNIVERSITY
IN KRAKÓW

Wrocław University
of Science and Technology

Stylometric Analysis of Large Language Model-Generated Commentaries in the Context of Medical Neuroscience

Jan K. Argasiński, Iwona Grabska-Gradzińska, Karol Przystalski, Jeremi K. Ochab and Tomasz Walkowiak

original papers

published commentaries

generated commentaries

NeuroImage 213 (2020) 116694

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

ELSEVIER

Establishing norms for error-related brain activity during the arrow Flanker task among young adults

Michael J
Annmari

[a] Department of
[b] Department of
[c] Department of

ARTICL

Keywords:
ERP
ERN
Pe
Flanker
Norms

CORTEX 164 (2023) 1–10

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

ELSEVIER

Research Report

The origin of pleasant sensations: Insight from direct electrical brain stimulation

Cécile Villard [a], Zoé Dary [b], Jacques Léonard [b], Samuel Medina Villalon [a,c], Romain Carron [d], Julia Makhalova [a,c], Stanislas Lagarde [a,c], Christophe Lopez [b] and Fabrice Bartolomei [a,c,]*

[a] *APHM, Timone Hospital, Epileptology Department, Marseille, France*
[b] *Aix Marseille*
[c] *Aix Marseille*
[d] *APHM, Timon*

ARTICLE

Article history:
Received 22 No
Reviewed 21 Fe
Revised 15 Mar
Accepted 28 Ma

INVITED REVIEW

Mapping the Unconscious Brain: Insights From Advanced Neuroimaging

Abid Y. Qureshi* and Robert D. Stevens†

*Department of Neurology, University of Kansas Medical Center, Kansas City, Missouri, U.S.A.; and †Departments of Anesthesiology and Critical Care, Neurology, Radiology, and Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, Maryland, U.S.A.

Summary: Recent advances in neuroimaging have been a preeminent factor in the scientific effort to unravel mechanisms of conscious awareness and the pathophysiology of disorders of consciousness. In the first part of this review, we selectively discuss operational models of consciousness, the biophysical signal that is measured using different imaging modalities, and knowledge on disorders of consciousness that has been gleaned with each neuroimaging modality. Techniques considered include diffusion-weighted imaging, diffusion tensor imaging, different types of nuclear medicine imaging, functional MRI, magnetoencephalography, and the combined transcranial magnetic stimulation-electroencephalography approach. In the second part of this article, we provide an overview of how advanced neuroimaging can be leveraged to support neurological prognostication, the use of machine learning to process high-dimensional imaging data, potential applications in clinical practice, and future directions.

Key Words: Disorders of consciousness, Neuroimaging, Positon emission tomography, Functional MRI, Transcranial magnetic stimulation, Magnetoencephalography.

(J Clin Neurophysiol 2022;39: 12–21)

Recent insights on coma and other disorders of consciousness (DOC) have been tightly coupled to advances in brain image acquisition and analysis. The aim of providing a mechanistic account of cognitive phenomena and specifically of consciousness, one of the highest aspirations in neuroscience, seems increasingly within the realm of scientific inquiry and measurement. The scale of measurement matters. It was unlikely that methods such as electrophysiologic cell recordings and tracer studies at the neuronal level could generate insights into the

OPERATIONAL MODELS OF CONSCIOUSNESS

Loss of consciousness can be decomposed to pathological changes involving two highly interconnected systems: arousal based in the ascending reticular activating system and awareness reflecting higher-order corticocortical and corticosubcortical circuits. Coma, an acute loss of both arousal and awareness, typically resolves in 4 weeks at which time the patient either

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

ELSEVIER

Commentary

A commentary on establishing norms for error-related brain activity during the arrow flanker task among young adults

Peter E. Clayson [a,*], Emily S. Kappenman [b], William J. Gehring [c], Gregory A. Miller [d,e], Michael J. Larson [f,g]

[a] *Department of Psychology, University of South Florida, Tampa, FL, USA*
[b] *Department of Psychology, San Diego State University, San Diego, CA, USA*
[c] *Department of Psychology, University of Michigan, Ann Arbor, MI, USA*
[d] *Department of Psychology, University of California Los Angeles, Los Angeles, CA, USA*
[e] *Department of Psychiatry and Biobehavioral Sciences, University of California Los Angeles, Los Angeles, CA, USA*
[f] *Department of Psychology, Brigham Young University, Provo, UT, USA*
[g] *Neuroscience Center, Brigham Young University, Provo, UT, USA*

ARTICLE IN

Keywords:
Standardization
Normative database
Error-related negativity (ER
Error positivity (Pe)
Event-related potentials (ER

We appreciate the
mative data for the e
ity (Pe) components o

CORTEX XXX (XXXX) XXX

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

ELSEVIER

Commentary

Positive emotions elicited by cortical and subcortical electrical stimulation: A commentary on Villard et al. (2023)

ROOF

Fausto Caruana

Institute of Neuroscience, National Research Council of Italy (CNR), Via Volturno 39/E, 43125 Parma, Italy

To investigate t
elicit emotions.
technical const
niques prevent

LETTERS TO THE EDITOR

Commentary on "Mapping the Unconscious Brain: Insights From Advanced Neuroimaging"

To the Editor:

We have read with attention and enthusiasm the article "Mapping the Unconscious Brain: Insights from Advanced Neuroimaging" by Qureshi and Stevens[1] which reviewed models of consciousness and the evidence related to neuroimaging as a prognostic tool for clinical practice. The authors describe neuroimaging and neurophysiology tools including structural and functional MRI, nuclear imaging, magnetoencephalography (MEG), and transcranial magnetic stimulation combined with electroencephalography (EEG) as approaches that may be useful in the study of disorders of consciousness.

This review provides a comprehensive and clinically useful guide to techni-

head within a helmet not covering the face, MEG is more accessible for patients than MRI. So, there should be little difficulty completing an MEG recording on a patient who is comatose. In fact, even those patients who are mechanically ventilated can have MEG successfully completed.[5]

The study of disorders of consciousness with advanced neuroimaging and neurophysiology tools has a promising future, and we would suggest that MEG could contribute unique information to further understand the mechanisms behind these important conditions.

Anto Bagić†
Susan Bowyer‡
Michael Funke§
Ismail Mohamed||
Jeffrey R. Tenney¶
Wenbo Zhang#
Andrew Zillgitt**
†University of Pittsburgh

children - an institution's experience. *Front Hum Neurosci* 2021;15:667777.

In Reply:

We read with considerable interest the commentary from Dr Bagic et al. We agree that magnetoencephalography (MEG) is a valuable modality which has the potential to enable critically important insights in the evaluation of patients with disorders of consciousness. The spatial resolution of MEG is an area of active investigation, with some studies suggesting millimeter resolution, higher than can be achieved with EEG albeit much lower than possible with MRI[1]

However, the research cited by Bagic et al. should be interpreted in context. When Stefan et al. state, "spatial resolution… can be as low as a few millimeters," they were citing the work of Oishi et al. In that work, Oishi et al. determined that at the frontal lobe, "epileptiform discharges extending over a 3-cm^2 area produced a strong enough extracranial magnetic field to be recorded by MEG," and in the basal temporal region, "epileptiform discharges needed to extend

Why would you do that?

except - it is interesting

for LLMs in medicine
the ability to **distinguish** IS important

integrity and trustworthiness of medical research and its implementation

high standard of **ethical transparency**

aids in the ongoing **evaluation** and improvement **of LLMs** themselves

Why commentaries?

they are part of **scientific ecosystem** but **do not require inherently new data**

How we generated the texts?

GPT - 4

Gemini

Med-PaLM 2

The prompt

1. **Paragraph**:

Given the following article, write a commentary article to be published in the same journal. Consider only the criticism of the methodology and the interpretation of the results. Do not summarise the whole text. Cite the scientific papers with your arguments.
Use only real, published scientific work:

2. **Citation of the original paper** including title and full journal name.

3. **Phrase**:

The original article is provided below:

4. **The text of the original research paper**
with abstract, highlights (when apply) etc. but without references.

evaluation methods

quantitative

qualitative

Evaluation methods (1) quantitative - R ‘*stylo*’

STYLO

Principal
Component
Analysis

of the covariance matrix
of the feature frequencies

Most **F**requent **W**ords

Culling 0-25%

Bootstrap
Consensus
Trees

Img source: https://serhack.me/articles/unveiling-anonymous-author-stylometry-techniques/

Evaluation methods (1) quantitative - *Jeremi Ochab & Tomasz Walkowiak own pipeline*

Ochab, J.K., Walkowiak, T.:
A pipeline for interpretable stylometric analysis.
In: Digital Humanities 2024: Conference Abstracts.
George Mason University (GMU), Washington, D.C. (2024)

Spacy ‘en_core_web_lg’

model for preprocessing steps (including tokenisation, named entity recognition, dependency parsing, and part-of-speech annotation)

tokenisation, **n**amed **e**ntity **r**ecognition, dependency parsing, part-of-speech annotation

LightGBM

as the state-of-the-art DART boosted trees classifier

SHAP

(SHapley Additive exPlanations) for computing explanations

Sci-Kit Learn

for feature counting and cross-validation

Img source: https://serhack.me/articles/unveiling-anonymous-author-stylometry-techniques/

Evaluation methods (2) qualitative

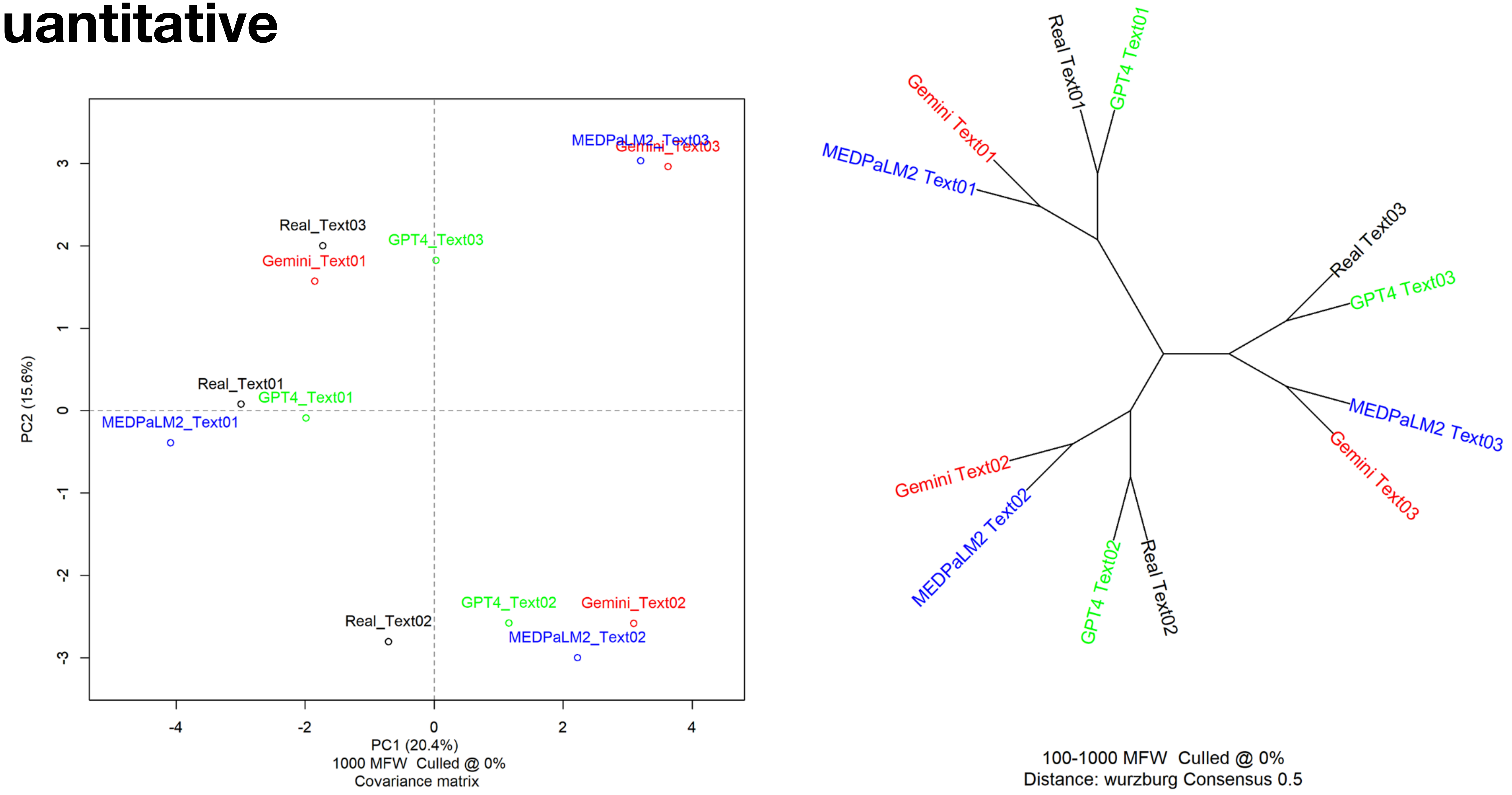
Annotation

1. accurate **summarization and referencing** of original research,
2. correct **references** to real academic papers,
3. proper **abstraction** of relevant knowledge from the cited papers,
4. coherent **argumentation** of presented arguments,
5. realistic **numerical results**, tables, or figures,
6. strict scientific knowledge – in terms of **factual correctness**,
7. strict scientific knowledge – in terms of being **state-of-the-art**,
8. **fitting structure**/argumentation as expected from a commentary,
9. **pertinent tone/style** as expected from a commentary,
10. **qualitatively new insight** with respect to the original paper.

The responses:
Yes, **No** or **Partly/Not applicable**

results

Results - Quantitative

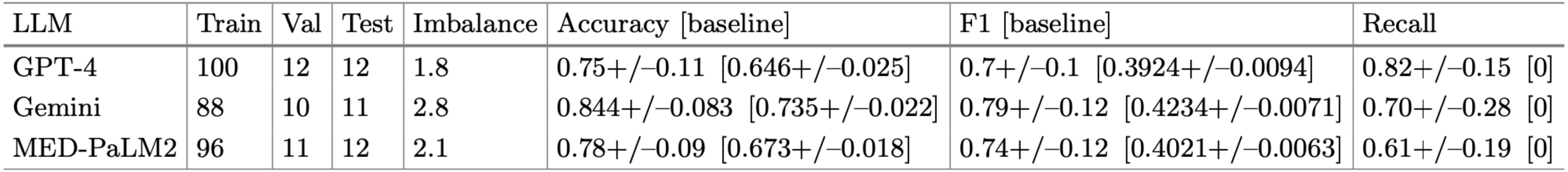


(left) Covariance PCA, (right) Bootstrap Consensus Tree

(i) the texts mostly cluster according to which paper they were commenting,

(ii)GPT-4's output consistently clusters with the real texts, while the other two models form separate clusters

Results - Quantitative



LLM	Train	Val	Test	Imbalance	Accuracy [baseline]	F1 [baseline]	Recall
GPT-4	100	12	12	1.8	0.75+/–0.11 [0.646+/–0.025]	0.7+/–0.1 [0.3924+/–0.0094]	0.82+/–0.15 [0]
Gemini	88	10	11	2.8	0.844+/–0.083 [0.735+/–0.022]	0.79+/–0.12 [0.4234+/–0.0071]	0.70+/–0.28 [0]
MED-PaLM2	96	11	12	2.1	0.78+/–0.09 [0.673+/–0.018]	0.74+/–0.12 [0.4021+/–0.0063]	0.61+/–0.19 [0]

LGBM classification results of 50-token samples.

The left-hand side of the table provides the median number of samples (across all cross-validation runs) in training, validation, and test sets and the ratio of the numbers of real to fake samples.

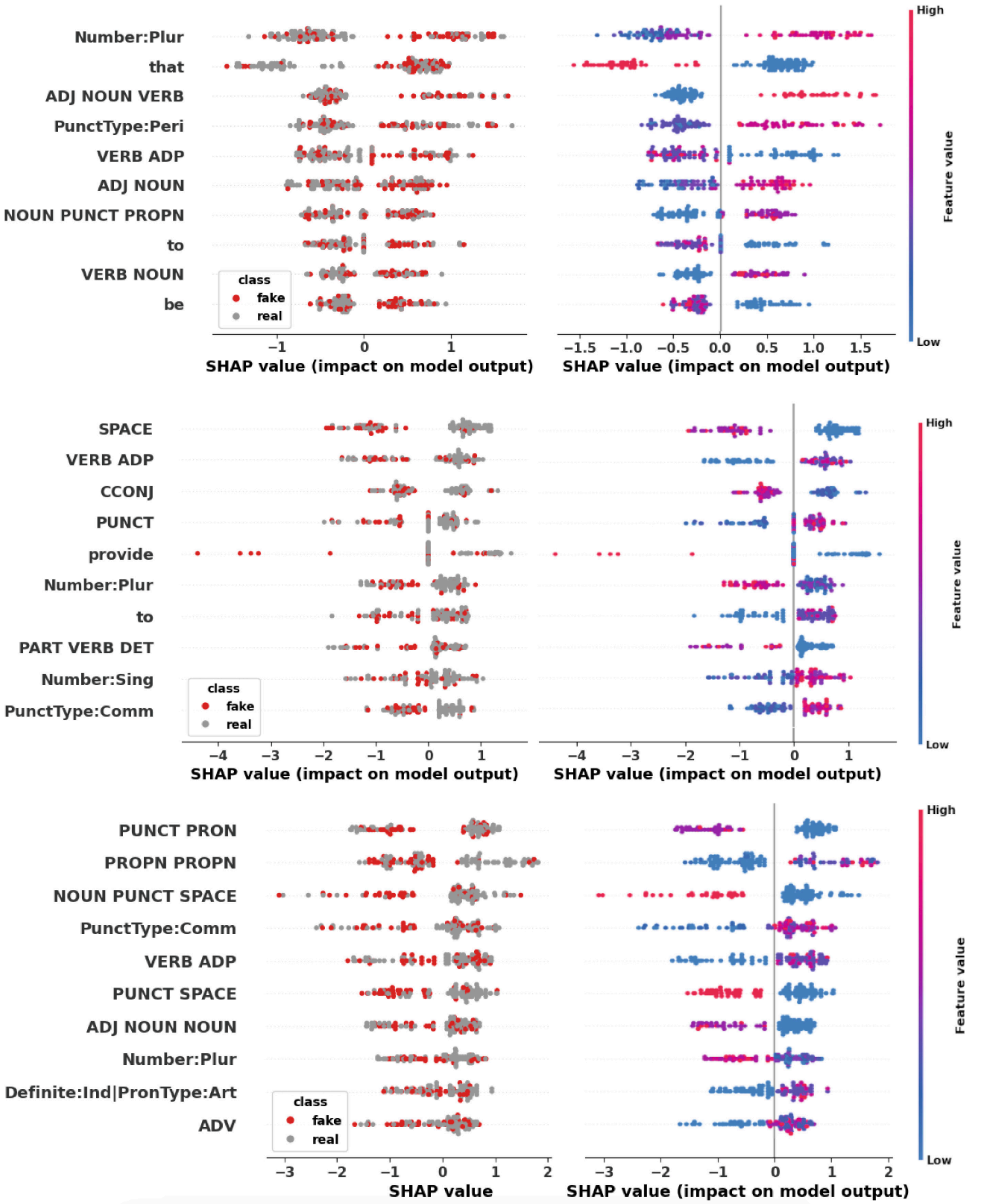
The performance metrics are provided against the baseline dummy classifier in square brackets.

Results - Quantitative

SHAP values
of the first 10 features
most important for classifying real commentary
vs (top) GPT-4,
(middle) Gemini,
and (bottom) MED-PaLM2.

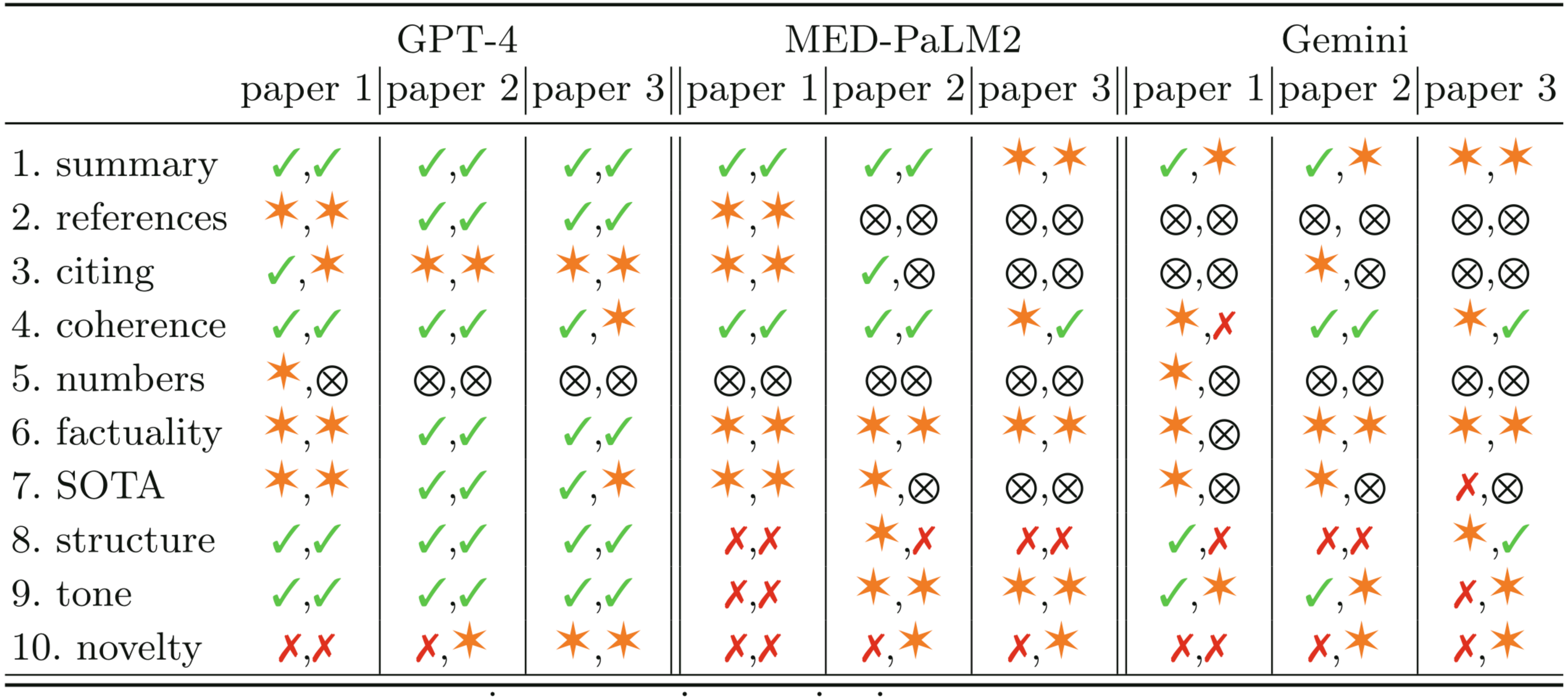
Each point is a 50-token text sample coloured
(left) by its class membership
(right) by its feature intensity.

Positive SHAPs point toward real texts,
and negative toward fake ones.



Results - Qualitative

Annotation



	GPT-4			MED-PaLM2			Gemini		
	paper 1	paper 2	paper 3	paper 1	paper 2	paper 3	paper 1	paper 2	paper 3
1. summary	✓,✓	✓,✓	✓,✓	✓,✓	✓,✓	✶,✶	✓,✶	✓,✶	✶,✶
2. references	✶,✶	✓,✓	✓,✓	✶,✶	⊗,⊗	⊗,⊗	⊗,⊗	⊗, ⊗	⊗,⊗
3. citing	✓,✶	✶,✶	✶,✶	✶,✶	✓,⊗	⊗,⊗	⊗,⊗	✶,⊗	⊗,⊗
4. coherence	✓,✓	✓,✓	✓,✶	✓,✓	✓,✓	✶,✓	✶,✗	✓,✓	✶,✓
5. numbers	✶,⊗	⊗,⊗	⊗,⊗	⊗,⊗	⊗⊗	⊗,⊗	✶,⊗	⊗,⊗	⊗,⊗
6. factuality	✶,✶	✓,✓	✓,✓	✶,✶	✶,✶	✶,✶	✶,⊗	✶,✶	✶,✶
7. SOTA	✶,✶	✓,✓	✓,✶	✶,✶	✶,⊗	⊗,⊗	✶,⊗	✶,⊗	✗,⊗
8. structure	✓,✓	✓,✓	✓,✓	✗,✗	✶,✗	✗,✗	✓,✗	✗,✗	✶,✓
9. tone	✓,✓	✓,✓	✓,✓	✗,✗	✶,✶	✶,✶	✓,✶	✓,✶	✗,✶
10. novelty	✗,✗	✗,✶	✶,✶	✗,✗	✗,✶	✗,✶	✗,✗	✗,✶	✗,✶

The responses:
Yes, **No**, **Partly**, **Not applicable**

The inter-annotator reliability was good as measured by ordinal Krippendorff's alpha, α = 0.77, 95% CI[0.67,0.86].

Conculsion

We demonstrate the **possibility of applying stylometric methods** for analyzing computer-generated texts in scientific domain.

Scientific and domain-specific texts are significantly more challenging to generate effectively due to their grounding in **real knowledge and facts, which cannot be easily summarized** from a general knowledge base.

These types of errors produced by the state-of-the-art language models can be assessed only by manual qualitative evaluation.

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

Wrocław University
of Science and Technology

CLARIN
Common Language Resources and
Technology Infrastructure

Thank you!

Acknowledgements. The publication was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857533. This publication is supported by Sano project carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund.

JKO and TW's research was financed by the European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19. JKO's research has been supported by a grant from the Priority Research Area DigiWorld under the Strategic Programme Excellence Initiative at Jagiellonian University.

European
Funds
Smart Growth

Republic
of Poland

European Union
European Regional
Development Fund

Jan K. Argasiński, Iwona Grabska-Gradzińska, Karol Przystalski, Jeremi K. Ochab and Tomasz Walkowiak

In our paper we **compare** artificially generated papers
with human-written scientific literature.

By matching LLM-produced text
with published commentaries on existing medical papers.