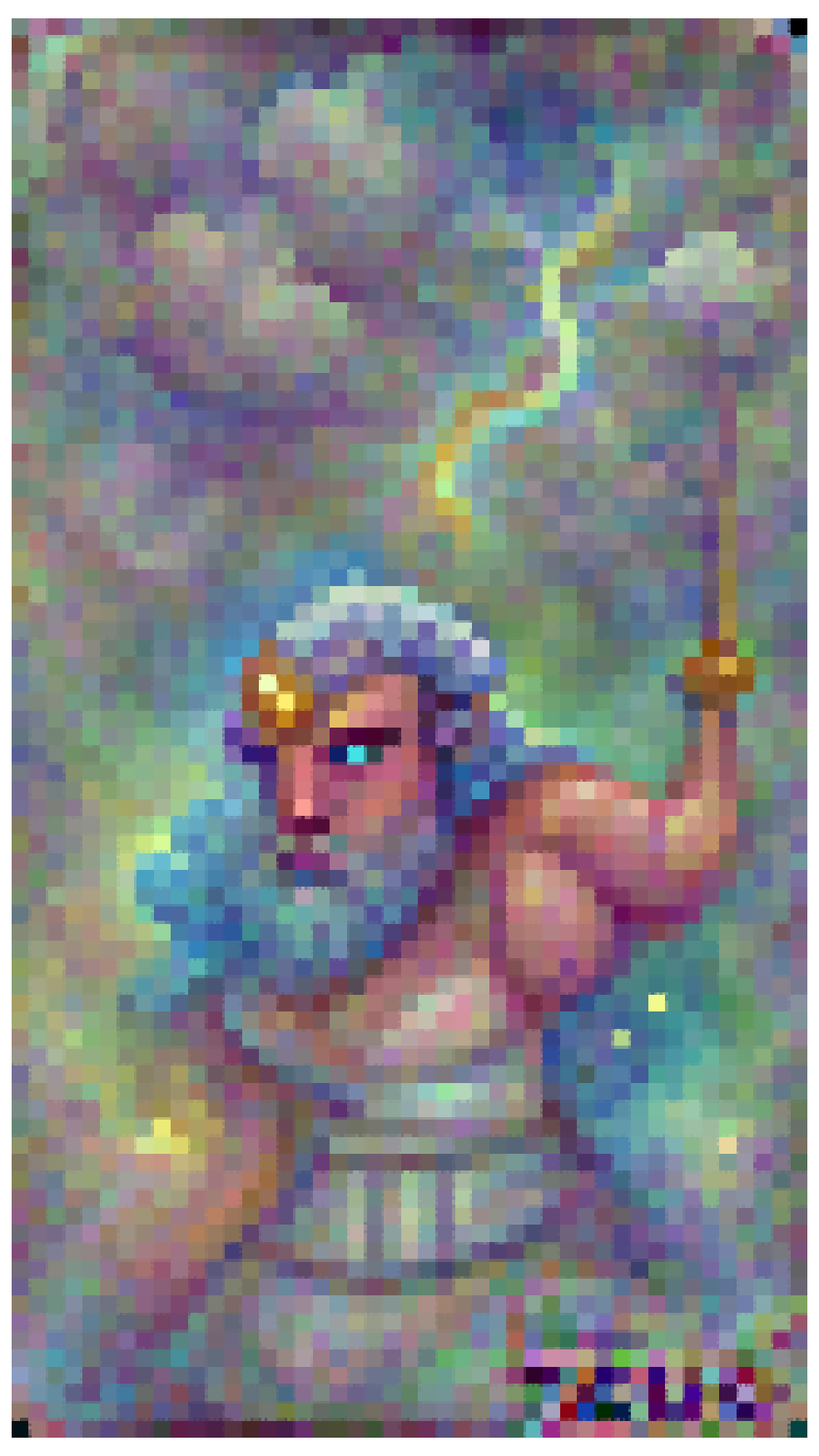

BERT in Plutarch's Shadows

Ivan P. Yamshchikov[1,2], Alexey Tikhonov[3], Yorgos Pantis[1],
Charlotte Schubert[4] and Jürgen Jost[1]

[1] Max Planck Institute for Mathematics, Leipzig; [2] CEMAPRE University of Lisbon; [3] Inworld.AI, Berlin, [4] University of Leipzig

- The extensive surviving corpus of the ancient scholar Plutarch of Chaeronea (ca. 45-120 CE) also contains several texts which did not originate with him and are attributed to an anonymous author Pseudo-Plutarch.
- These works are Placita Philosophorum, De Musica and De Fluviis;
- This paper presents a BERT language model for Ancient Greek. The model discovers previously unknown statistical properties relevant to these literary, philosophical, and historical problems and can shed new light on this authorship question;
- In particular, the Placita Philosophorum, together with one of the other Pseudo-Plutarch texts, shows similarities with the texts written by authors from an Alexandrian context (2nd/3rd century CE).

	Symbols per Token		Words per Token	
Tokenizer	Greek BERT	Multilingual BERT	Greek BERT	Multilingual BERT
Modern Greek	4.52	2.55	0.72	0.41
Ancient Greek	2.98	1.9	0.46	0.31

Table 1: In comparison with multilingual BERT, Greek BERT tokenizer shows a higher number of symbols and words per token for both Modern and Ancient Greek

	Validation accuracy
Greek BERT	**80%**
Greek BERT no MLM-transfer	78%
Multilingual BERT	78%
Naive Bayes Classifier	43%
Random authorship attribution	6%

Table 2: After MLM training and ten epoch of fine-tuning for authorship attribution, the validation accuracy of Modern Greek BERT is slightly higher than that of the Multilingual BERT after similar fine-tuning procedures. Modern Greek BERT fine-tuned for authorship attribution without MLM transfer learning phase shows lower validation accuracy. All BERT-based classifiers significantly outperform the Naive Bayes Classifier that uses the two thousand most frequent unigrams. Another baseline attributes one of seventeen labels to the text at random.

	G	O	P	CD	FJ	PJ	A	CP	AA	S	L	CA	Ap	P	SE	DC	other
Galenus	**416**	7	5	1	1	4	5	5	6	3	3	5	1	0	6	10	22
Origenes	2	**396**	0	1	6	4	5	12	1	3	0	24	0	0	6	5	35
Plutarchus	6	3	**390**	3	9	8	17	2	2	5	2	5	12	1	6	13	16
Cassius Dio	1	0	8	**428**	5	2	2	0	7	8	2	1	17	6	0	7	6
Flavius Josephus	3	3	10	5	**418**	2	4	6	8	9	6	1	8	0	4	9	4
Philo Judaeus	5	10	13	3	16	**403**	3	3	2	3	3	12	0	0	11	8	5
Athenaeus	11	6	17	4	4	2	**368**	4	7	11	9	7	2	6	6	14	22
Claudius Ptolemaeus	3	0	0	0	0	1	0	**480**	0	8	0	0	0	0	5	0	3
Aelius Aristides	7	6	6	6	7	2	5	0	**368**	8	10	6	1	3	3	40	22
Strabo	4	5	9	0	3	2	7	1	9	**432**	4	1	3	6	4	4	6
Lucianus	2	3	6	1	5	4	9	0	13	9	**360**	12	5	6	6	30	29
Clemens Alexandrinus	8	28	3	4	10	14	4	1	6	5	8	**349**	0	5	17	11	27
Appianus	1	0	10	18	8	2	2	1	3	2	5	3	**437**	0	0	3	5
Pausanias	0	1	1	2	0	0	2	0	4	3	2	3	2	**472**	0	3	5
Sextus Empiricus	2	4	6	0	1	1	4	1	2	2	1	11	0	0	**446**	7	12
Dio Chrysostomus	2	4	12	9	5	3	7	0	9	10	10	9	6	4	2	**398**	10
other	17	23	22	7	6	15	32	14	10	6	12	18	6	9	40	21	**242**

Table 3: The confusion matrix of the obtained authorship classifier. Every horizontal line sums up to 500 sentences by the corresponding author that were set aside for validation. Every column shows the number of sentences labelled by classifier as sentences authored by the corresponding author.

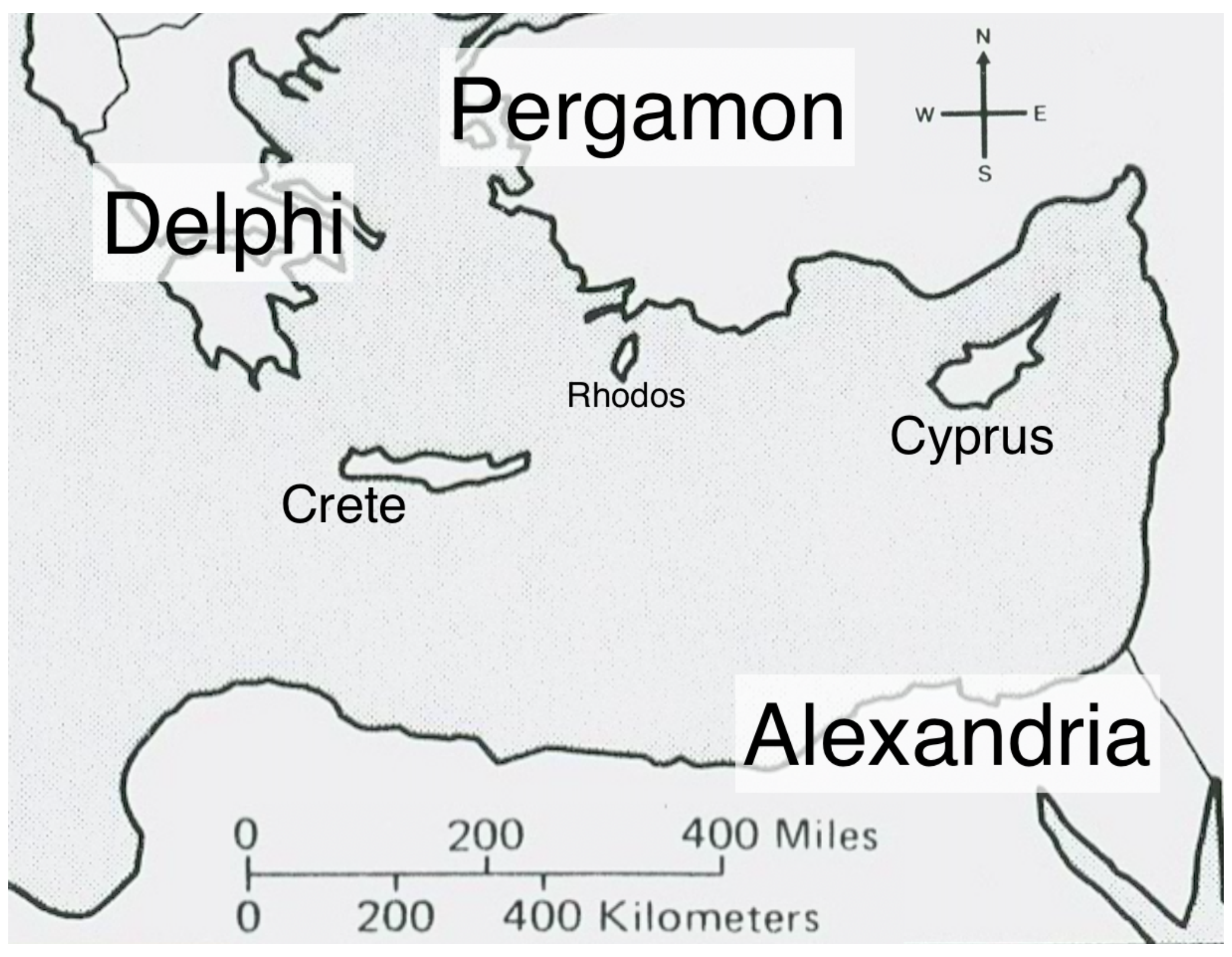

Figure 1: A map showing relative position on three potential regions relevant for authorship attribution of Pseudo-Plutarch documents.

Predicted Region	Pergamon Region	Alexandria Region	Delphi Region	Other Regions
Pergamon	**83%**	3%	3%	7%
Alexandria	5%	**77%**	7 %	10%
Delphi	4%	5%	**81%**	8%
Other	8%	15%	9%	**75%**

Table 4: Results of the BERT-based regional classifier on 4000 sentences set aside for validation.

	Sample Size	Top 1	Top 1 Share	Top 2	Top 2 Share	Top 3	Top 3 Share
De Fluviis	310	Athenaeus	22%	Others	21%	Strabo	19%
De Musica	285	Athenaeus	21%	Plutarch	18%	Sextus Empiricus	14%
Placita Philosophorum	928	Others	36%	Claudius Ptolemaeus	20%	Sextus Empiricus	11 %

Table 5: The most frequently attributed authors in the three Pseudo-Plutarchean texts.

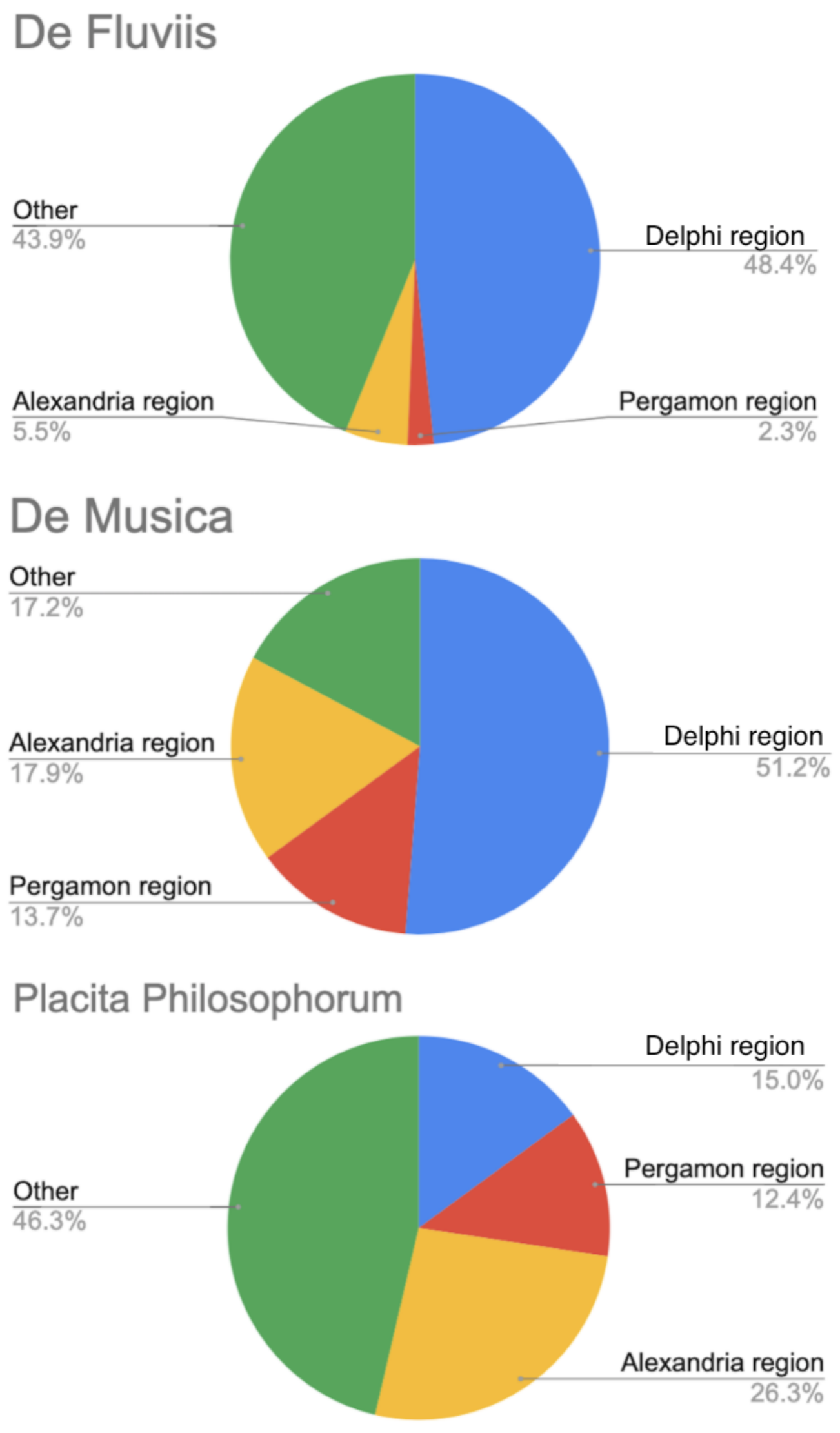

Figure 2: All three Pseudo-Plurach documents show significantly different percentages of sentences attributed to a certain region. In particular, Placita Philosophorum is the only document where Delphi is not a dominant region, while Alexandria is the most frequent identifiable region.

https://huggingface.co/altsoph/bert-base-ancientgreek-uncased