Revision round #1

Decision for round #1 : *Revision needed*

Dear all,

Please see the important recommendations of both reviewers, especially regarding the exposition of the model you intend to show. Some other aspects related to data visualization seem equally pertinent to make the best use of your narrative.
*by **Daniel Carvalho**, 22 Dec 2024 08:54*
Manuscript: **https://doi.org/10.5281/zenodo.13283972**
version: 1

Thank you for the helpful reviews, we have adapted the document accordingly. The comments in the pdf largely overlapped with the comments below, and for those which did not overlap we addressed them in our new version as well. Below in blue we mention for each comment how we addressed it.

Review by Simon Carrignon, 24 Sep 2024 17:15

The paper propose an interesting approach that uses large language model to detect archaeological sites from a specific culture that that may have been missed in previously build database using archaeological reports. The method rely on searching for specific terms in publicly available pdfs, and use of LLMs to ensure the hits are well charactised and not false positives.

ALthough I do thinkg the paper is interesting and present an original and well caried piece of work, there is still a few points I would like the authors to answer.

First of all, it is still not very clear to me the relationship between AGNES, NER, BERT, and the relevance classifications presented in Tables 2 to 4. What is BERT bringing exactly? When given the search terms in Table 1 and thanks to its training, is it able to classify reports that fall within “Vlaardingen Groep,” (for example) even if the terms are not necessarily mentioned or not in this exact order or spelling in the report? Then, how are the reports classified as relevant or not? Is that done by BERT too, or is it manually checked?

This is very likely due to my lack of knowledge of the model used, and of the previous papers cited by the authors, but I am struggling to understand how this is different from simply looking for the term (with some regex) and then manually check the reports returned. I do trust the authors that this is indeed different, but I struggle to see how exactly this is done, and I think the authors could clarify this with one or two sentences.

Thank you for your suggestion. As this is not clear to the reader we now added the following sentences to paragraph 2.2:

“Following the methodology presented by Brandsen and Lippok we manually checked the hits.”

“It is worth noting here that the free text query in Table 1 is simply doing a term match in ElasticSearch, while the start date and end date are making use of the time periods detected by BERT. For this particular study, we do not make use of the other entity types, such as artefacts or materials.”

The web interface of Agnes is also briefly mentionned, and when digging the supplementary material, one can find the IP address to interact with the logiciel. And again, this may be due to my own ignorance about the project and previous publication, but why no address to access AGNESS is given yet? is it because the project is still in developpement? Also, is the code behind AGNES, and the interactions with BERT and the NER methods going to be published? The overal architecture of the project seems to be an amazing achievement, that could be used by any country/institutoin with publicly available archaeological reports in pdf, is there any plan to publish/share these tools? I understand they may not be ready but maybe the authors could mentionned it in the paper.

This is a very good point, it was an oversight to not include the link to AGNES, we have now added that in section 2.1:

“The system can be freely accessed via https://agnessearch.nl, and all code/models will be made available open access at the end of the AGNES project.”

I am not too sure about the network visualization of the results. They may be a nice alternative to simple barplot with the percentage per query, but then the authors should say a sentence or two justifying this choice and explaining a bit more what do these network tells us.
To follow with these network: what is the difference between multiple edges (like between "Vlaardingen Stein" and "not relevant") vs thicker edges (like between "Vlaardingen Stein Wartberg" and "semi-relevant")? Also, wouldn't some normalization of the edge/node sizes help visualize the data? such as taking the log or the square root of the values?

We agree that the multiple edges creates confusion, especially because with many edges these were now visualed as single thick edges (when in reality these consist of multiple edges). Therefore, we now ranked the edges by weight to improve the visualization of the data.

To accommodate the first part of the comment we added a few sentences to our methodological section as justification for our visualisation methods:

“The data is visualised in distribution maps in QGIS to highlight the spatial distribution of our findings. To visualise the relationships in our data we used network visualisations which were created in Visone. We adopted network graphs because these provide a visual representation in which the results of multiple queries can be summarised in a single relational graph.”

Finally: I very much appreciated the discussions about the discrepancy between the results of this paper and the known literature, and how the authors interprate it. This also make a good case on how AI can help understand how history and sociology of a discipline can bias th results. I found the maps particularly interesting in that regards, especially the one that shows how the culture is described in different reports, revealing non-random customs among the archaeologists excavating these sites. Nonetheless, I think these maps could be better rendered. The choice of color and the large size of the dots used to represent the sites make it very hard to see clear patterns. Do the geology maps really need to be included? maybe only some DEM would be enough? Many of the colors between the dots and the base map are very similar, making it difficult to interpret.

Thank you for the comment, following this we removed the Palaeographic map and adopted a more neutral map which indeed makes it easier to read. We also downscaled the dots as suggested by the reviewer. We did not include a DEM because the western Netherlands are virtually flat and we feel that elevation is an insignificant variable for this area. We did keep major rivers in the visualisation as those are important in relation to the landscape zones which were inhabited in prehistoric times.

Using smaller dots to represent the sites, on a map with fewer colors —perhaps even including only rivers and the sea, if slopes still makes it too messy?- could allow to remove the 'zoomed' versions of the maps? Then the two figures could be joined in a single one with two panels: previously found/not found on one side and cultural attribution on the other? Pushing it further: as there are only four categories for found/not found, different shapes of circles could be used to encode this, allowing to have everything on one single combined map? However, I understand this might make it less readable. Maybe the different shape could then be combined with the year of the excavation, that may illustrate nicely the fact that older report are not in pdf and show how the missing sites are the one published long time ago

As suggested we indeed removed the Palaeographic map with different colours in favour of a neutral map, only including river and sea data to prevent messiness. Nevertheless, even with small dots the dense clustering in the Nijmegen area makes this part of the map difficult to read. Therefore we did keep the submap, and spacewise we would still need the two maps. The possibility of combining the two maps would make it less readable hence we prefer to keep this as two separate figures. The inclusion of a year of excavation would be problematic because excavations often consist of multi-year campaigns. Therefore, we feel that our observation about older excavations would not be accurately represented in such a map. For example, the site Zandwerven was excavated in 1929, this is the earliest excavation of a VLC site. Nevertheless, the site is also found in AGNES. This is because the site has been subsequently excavated until the 1980’s, and it was also included in a 2001 report on Neolithic sites in the West Frisia area. Adding the excavation year (1929) would cause confusion regarding our general observation that a lot of the sites which are not found in AGNES consist of older excavations. We feel that our in-text discussion with examples, for example on the sites in the Molenlanden, better illustrates this point we intended to make.

Minor remark:

- What does "OG", in the sentence " the OG Large Language Model (Devlin et al. 2019)" means? as it's not defined before i read it as Original Gangster, but I doupt the authors wanted to use this temrinology?

To accommodate this we replaced OG with the more neutral term “original”.

- The output of the csv supplemtary material of the table 2 as like 1000 row but only 167 sites have something, is that

Thank you for noticing this, we have changed this in the new version.

Overall this is an interesting paper, that showcase a nice use of AI in archaeology, and I think it would fit well PCI Archaeology. Modulo the comment I make here I think it will be a great publication!

Review by anonymous reviewer 1, 25 Nov 2024 19:05

Disclaimer

I am a geographer and archaeologist trained in new archaeological methods, and I have worked on machine-learning applications in archaeology for a few years. However, I am neither a Dutch archaeologist specialist nor a computer scientist. Therefore, my knowledge of Dutch archaeological cultural problematics and large language model hyper-tuning is limited. I am also a non-native English speaker.

Title and abstract:

Major issue:

- *Does the title clearly reflect the content of the article? [ ] Yes, [x] No (please explain), [ ] I don't know* “From the River to the see” sentence in the title does not fit any of the elements discussed in this pre-print. If I agree with using a “pre-title” or a short sentence that might attract the reader here, it does not seem to correlate with the rest of the subject. One major problem is using the Large Language Model word to describe BERT (Cf. below).

Thank you for your comment. We have removed the pre-title “From the River to the Sea” from our title. However, we disagree that the term Large Language Model cannot be used to describe BERT. BERT is commonly acknowledged as being a Large Language Model. See for example the publication by Bommasani et al. 2022 (Bommasani et al. 2022, 158):

“Large language models such as BERT and M-BERT are capable of cross-lingual transfer, which — when the models are open-sourced — may allow for adaptation to languages which otherwise would have too few texts available [Wu and Dredze 2019; Wang et al. 2020a].”

Bommasani, R., et al. (2021). On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

Or the publication by Banerjee et al. 2025 (2025, 1):
“This study investigates the internal mechanisms of BERT, a transformer-based large language model, with a focus on its ability to cluster narrative content and authorial style across its layers”

Banjeree, A., et al. (2025). Exploring Narrative Clustering in Large Language Models: A Layerwise Analysis of BERT. arXiv *preprint arXiv:2501.08053*.

It is unfortunate that our use of this term is now erroneously cited as a major issue when in fact our use of the term Large Langauge Model, to refer to BERT, is appropriate. But since BERT was only a minor part of the study we have removed the term Large Language Models from the title as suggested.

- *Does the abstract present the main findings of the study? [ ] Yes, [x] No (please explain), [ ] I don’t know.* The abstract is clear for most of the points except for the additional goal of the collected results besides testing the model. In the discussion, several points are debated on the VLC that are not mentioned in the abstract. I would recommend adding one or two short sentences on these cultural interpretations.

  Indeed we agree that the discussion on cultural interpretations became an important part of the study, which warrants a mention in the abstract. Therefore, we added the following sentences here for clarification:

  “Finally, our study highlighted the fact that clear terminology to describe Vlaardingen Culture sites is presently lacking. As such the study provided interesting insights into the terminologies employed in development-led archaeology.”

Minor issue:

- **Line 18** There is a lack of clarity in describing the Vlaardingen Culture characteristics.

  While we fully agree that the Vlaardingen Culture should be better defined in the article (in line with the comment below under “Major Issue”) we do not feel that this should be done in the abstract. It is not common practice to define archaeological cultures in abstracts, unless the primary goal of an article is to define such cultures (which is not the case here). Therefore we feel that this description is best given in the introduction, rather than the abstract (which we now did).

- **Lines 38 – 39** The keywords use words similar to those in the title. Alternative ones for a broader range of results in ulterior referencing online would be suggested.

  Thank you for your helpful suggestion, we agree that the use of different keywords would help others to find the article online. Hence we have made the following changes to those keywords which were mentioned in the new title:

  Neolithic → Middle Neolithic;
  Vlaardingen Culture → Late Neolithic;

Introduction:

Major issue:

- *Are the research questions/hypotheses/predictions clearly presented? [] Yes, [x] No (please explain), [ ] I don't know.* As a non-expert in Dutch archaeology, my comment might be inaccurate. However, I would suggest a deeper explanation of the VLC culture, which is described here as "The vast majority of Vlaardingen Culture sites consist of artefact scatters without clear house plans (Van Gijn & Bakker 2005). Only a few sites contain clearly discernible house plans (Stokkel 2017; Van Beek 1990; Van Kampen 2013; Van Zoolingen 2021; Verhart 1992)." (**l.63-66**).

Thank you for your valuable comment. We agree that our initial description was insufficient as we did not accurately describe the subsistence strategies or material culture which is characteristic for the Vlaardingen Culture. To address this comment we added the following:

"In terms of subsistence strategies Vlaardingen Culture sites characteristically yield evidence for a mixed strategy which involves cereal cultivation, animal husbandry, as well as hunting, fishing and gathering (Raemaekers 2005; Van Gijn and Bakker 2005). The sites are predominantly located along the coastal dunes of the western Netherlands and along the major rivers in the western and central Netherlands (Raemaekers 2003). The ceramic assemblages consist mostly of quartz tempered undecorated S-shaped pots, occasionally with a row of perforations under the rim. Furthermore, clay discs and collared flasks are a regular occurance. The lithic assemblages are dominated by simple 'ad hoc' flake technologies. Flint axes, often found in a broken state, consist of oval axes of the "Buren-type". Arrowheads predominantly consist of transverse arrowheads, tanged points, and leaf-shaped arrowheads (Van Gijn 2010; Van Gijn and Bakker

2005; Van Regteren Altena et al. 1962c)."

- *Does the introduction build on relevant research in the field? [x] Yes, [ ] No (please explain), [ ] I don't know*

Structure:

The introduction's structure would need major revision, as the description of the objective (**lines 59 – 60**) comes before the description of the VLC and AGNES. This paragraph of lines **59 – 70** would need to be set at the end of the introduction or rewritten entirely so it does not mention elements that have not been introduced before.

Thank you for your suggestion, we have now moved the aims to the end, just before the research questions. In the first paragraph we introduced the AGNES search engine, and in the second we introduced the Vlaardingen Culture. Therefore, the aims now come after the introduction of these concepts.

For the whole 1.2 part (**lines 111 – 124**) I would suggest instead of the paragraph a table synthetizing all the information and including **lines 121 – 124** in the method part or legend of the table.

Thank you for your comment, however we feel that having the AGNES dataset as a separate paragraph helps to highlight the different datasets employed in this study. Now we have a clear separation between a) the Vlaardingen dataset presented in paragraph 2.1 and b) the AGNES dataset presented in paragraph 2.2. under a common heading 2 Data.

I would appreciate a more detailed paragraph on the recent influence of large language models in general and the explosion of transformers after Vaswani et al. 2017 paper.

We have added a couple of sentences about this topic in section 2.1

Language:

The English is good, but it is not easy to read. It would win in being read by a native English speaker

The text has been proofread, and some minor changes have been made throughout the paper.

- **Lines 47 – 49** The sentence is unclear, especially the "the end users" meaning.

  We updated the sentence to make it clear what we mean by end users (archaeologists searching for data)

- **Line 56**: Present first the full name of the tool and then the acronym in parenthesis, not the other way around.

  Thank you for your suggestion, we have changed the order.

- **Lines 72 – 77** The paragraph on Brandsen and Lippok's previous study, 2021, is unclear as it repeats itself.

  Thank you for your comment, but it remains unclear to us where this paragraph repeats itself. Perhaps the confusion lies in the last sentence, maybe it is unclear here to what study that sentence refers to. To clarify we have changed the 'in this study' to 'in the present study' to prevent any confusion.

- **Line 116** A general comment: All the prompts should be written in their native language (Dutch) in italics and then in English, like "*Archeologie*/Archaeology."

  Thank you for your suggestion, we assume that the reader here refers to our queries, not prompts. While we agree that having an English translation accompanying Dutch words might be useful we feel that the proposed format here will create significant confusion as every Dutch word in the text is either identical to its English translation (Vlaardingen/Vlaardingen), or it is nearly identical (Vlaardingen Cultuur/Vlaardingen Culture). Furthermore, all queries are clearly translated in table 1. Nevertheless, to accommodate this comment we now adopted the format previously used in the publication by Brandsen and Lippok. Here the original Dutch terms were indeed written in italics and they were accompanied by an English translation between brackets: '"*vlaardingen cultuur*" (Vlaardingen Culture)'. We feel that this should be sufficient to clarify the Dutch terms for any non-Dutch readers. We did however not translate identical terms as this would only create confusion.

Content:

On the general content of the introduction, one primary piece of information is missing: a clear description of the Vaardingen Culture. Only one sentence "The vast majority of Vlaardingen Culture sites consist of artefact scatters without clear house plans" (l.63-65) is giving information on this culture. However, a proper description of a chrono-cultural entity should include more information. How are we differentiating this culture from the other at the same chronological time (e.g. ceramic, metalwork, buildings)? What are its specificities?

We fully agree that we should have included more specifics. This comment is similar to the comment before where the reviewer also addresses the lack of a proper description of the Vlaardingen Culture. This issue is therefore addressed above.

Minor issue:

- **Line 63** There is no reference for the VLC chrono-cultural period.

  Thank you for this comment, we added the period behind the first mention of the Vlaardingen Culture in line 62 (removing it from line 65) to accommodate this comment.

- **Lines 96 – 99** No map is provided for representing this new area of excavation, or the "potential" extent of the VLC.

  Thank you for this suggestion. We agree that a map with the extent of the Vlaardingen Culture would be helpful for readers unfamiliar with this group. We similarly felt that, due to our later discussion which also involved the borderlands between the Stein group and the Vlaardingen Culture, it would be helpful to also include the distribution of the Stein group here. We added this as a new figure (1).

- **Lines 116, 118-119** It would be needed to provide a reference or hyperlink for the mentioned database/portals.

  Thank you for your comment, we now added links to the three portals.

Materials and methods:

Major issue:

- *Are the methods and analyses sufficiently detailed to allow replication by other researchers? [ ] Yes, [x] No (please explain), [ ] I don't know.* The main goal of this preprint is to provide a detailed process of the AGNES tool. However, if the reference to the Brandsen 2023 BERT model is present, no detailed information is furnished on the preprocessing of the data set or the model.

  Thank you for your comment, however, as stated in the abstract as well as elsewhere in the article the above is not the main goal of the paper. In the abstract we clearly specified the goals of the paper as follows:

  "The aims of this paper are twofold: 1) to provide an up-to-date overview of Vlaardingen Culture sites; 2) to evaluate the performance of AGNES in searching for period specific sites."

  We repeated these aims clearly in the introduction:

“The aim of this study is twofold: 1) we aim to provide an up-to-date overview of Vlaardingen Culture (3400-2500 BCE) sites; 2) we aim to evaluate the performance of AGNES in searching for period specific sites.”

We fail to see how the reader could have gotten from this paper that the overall aim was to provide a detailed process of the AGNES tool. Especially, considering we have aptly cited the Brandsen 2021 paper which specifically aimed to do this:

“**Brandsen, A.** 2021. *Digging in Documents - Using Text Mining to Access the Hidden Knowledge in Dutch Archaeological Excavation Reports.* PhD Thesis. Leiden University.”

We believe that the confusion regarding this issue lies with the reviewer and we don’t see how can be more clear as a third repetition in which we would again explicitly repeat our twofold aim would be unnecessary overkill.

- *Are the methods and statistical analyses appropriate and well described? [ ] Yes, [x] No (please explain), [ ] I don’t know.* There seems to be confusion about the Large Language Model definition.

  As addressed earlier in the review, our use of the term Large Language Model when referring to BERT is appropriate and in line with the use by other scholars (as cited in our earlier comment).

Structure:

The general structure of the chapter is clear except for the 2.2.1 part, which stands alone without any reason. Maybe it would need to include paragraph **lines 202 – 208** as part of the 2.2 part only. As it mentions the relevant table, it should be placed above **line 190.**

Thank you for your comment, indeed this is a very short section which maybe does not really warrant to stand alone. Indeed, it ties into the section before, as well as to the table presented above. However, we do feel that it fits after the description of the relevant, semi-relevant, and irrelevant categories. We begin this section with “Above we presented different categories of relevant, semi-relevant, and irrelevant hits.” which in our view highlights this logic. We do understand that this might however be clearer if we remove the heading for paragraph 2.2.1 hence we removed this heading so the section becomes part of paragraph 2.2. (now 3.2)

Language:

Similar general considerations as for the Introduction part stand here (Cf. above):

- **Lines 132 – 133** This sentence seems unnecessary.

  Thank you for your suggestion. The sentence referred to here is the following: *"Think of an excavation of a Roman encampment; the metadata is not going to mention a single Neolithic find (by-catch), but this find is only mentioned in the excavation report."*

  This sentence was added here to further clarify the difference between simply searching for a term (for example in Dans or ARCHIS) and the search in AGNES. In a comment the first reviewer stated "but I am struggling to understand how this is different from simply looking for the term". Hence we feel that this additional clarification is important as otherwise it might not be clear enough for some readers that this kind of full text search is of added value as opposed to the traditional metadata searches.

- **Lines 136 – 139** The sentence is unclear and needs re-writing.

  Thank you, but based on this remark it is not clear to us what is deemed to be unclear here. Therefore, except for the change from mediaeval to medieval we have not altered the sentence.

- **Line 137** The spelling of "medieaval" while correct for British writing only, is not commonly used. Please consider writing "medieval" without an "a" if submitting to an American journal afterwards.

  Thank you for your comment. If we would submit to an American journal we would indeed need to change British spelling words to American spelling. To accommodate the reviewer we have changed 'mediaeval' to 'medieval'.

- **Lines 137 – 138** With the same considerations as in the previous part, please include the prompts' original language (Dutch) in italics.

  Thank you for your suggestion. We have changed the medieaval to medieval as suggested. In these examples we refer to example queries (which is probably what the reviewer refers to when referring to prompts). In this sentence the following '('Flint' being both a material and a surname)' thus does not refer to a Dutch query 'vuursteen' it simply aims to show how the word Flint can be both a material and a surname. Hence in this sentence we feel that a Dutch translation is not warranted.

- **Line 146** The same consideration is given to the prompt and the "mediaeval" writing.

  Thank you for your suggestion. We have changed the medieaval to medieval as suggested.

- **Line 168** Table 1 The English translation does not need capital letters.

  Thank you for your suggestion, we have removed the capital letters from the English translations in table 1.

- **Line 183** Unclear sentence.

  Thank you for your comment. This sentence: "This was often the case with research plans." refers to how the above was often the case with research plans. To clarify this we have changed it to "Research plans for example often mentioned the Vlaardingen Culture, or Vlaardingen Culture sites when describing the archaeological potential of a study area. Such hits were therefore considered to be semi-relevant hits."

Content:

Enhanced the choice of BERT as in Brandsen 2023 "However, for specific domains and low-resource languages, BERT can still outperform LLMs."

A critical point is the confusion between LLMs and the BERT model, as expressed in " The NER is done using BERT (Bidirectional Encoder Representations from Transformers), the OG Large Language Model (Devlin et al. 2019). Similar to the newer GPT (Generative Pre-trained Transformer) models, BERT uses large amounts of unlabelled text data to pre-train a model, gaining an understanding of words and their contexts" (**l.149 – 152**). I do not know if the authors refer to an "OG" model, which would be an LLM, in which case you will need to give more precisions and a reference, or if they refer to BERT as an LLM. This would be inaccurate in the second case as the BERT does not have two main characteristics of LLMs. First, the number of parameters of BERT is "only" hundreds of millions, while the number of GPT3 or PALM parameters is hundreds of billions. Second, the BERT does not have a decoder phase in its architecture, contrary to GPT3 or other LLMs (Rogers et al. 2020). This confusion would need to be clarified as many parts of the pre-print mention BERT as an LLM would need to be reformulated.

While GPT3 and similar generative models are indeed bigger than BERT, BERT definitely is an LLM, and is consistently referred to as an LLM in the literature. LLMs are defined as 'large' models, meaning they use deep learning, are trained with self-supervised learning on huge datasets, and can solve NLP tasks. BERT aligns fully aligns with this definition. Whether a model has a decoder phase is irrelevant to the classification of an LLM, as LLMs can be either generative or classification models (or even just vector generators). While of course the term LLM is often used as a synonym for GPT like models in the media at the moment, it would be unjust to not class BERT as an LLM, nor its predecessors such as LSTMs.

The details of the pre-process are not given, which is another major issue. The paper would need to refer to the pre-training of the BERT model, both original and modified versions trained on the 60k documents of the DANS (Brandsen, 2023). Information on which BERT model is used (large or base) would be needed. The pre-treatment of a text (e.g. deleting of white space, common words, numbers) needs to be transparent to be replicable, and no information is given here.

The details of the pre-processing are not given, as no pre-processing needs to be done for BERT (and most other LLMs), and as such we did not do any pre-processing. These models work on plain, unprocessed text.

__Minor issue:__

- **Lines 165 – 166** There is no reference for the VLC chrono-cultural period.

  Thank you for your suggestion, we have now added the following reference to clarify this: (Raemaekers 2005, 271).

- **Lines 168** Table 1 would need to have combined rows for similar English translations.

  Thank you for your suggestion, however we feel that this is a matter of taste. In our setup we simply followed the setup of the table provided by Brandsen and Lippok 2021 (table1). We feel that having two separate columns here is relevant because the free text query is different from the translation in the sense that this part is entered into the AGNES search engine. Also because occasionally the query and translation are the same (Vlaardingen and Vlaardingen for example) we feel that having these in a single column will create confusion.

- **Lines 190 – 192** If categories 10 – 12 are mentioned, the mention of categories 1 – 9, related to a previous typology from Brandsen and Lippok, 2021, would be needed.

  Thank you for your comment, to clarify this we have included (1-9) in the sentence to clarify that those categories were derived from the previous publication:

“Irrelevant hits (1-9) were classified along the typology presented by Brandsen and Lippok (Brandsen & Lippok 2021).”

Results:

Major issue:

- *In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? [x] Yes, [ ] No (please explain), [ ] I don’t know*
- *Are the results described and interpreted correctly? [x] Yes, [ ] No (please explain), [ ] I don’t know*

Structure:

No comment on the structure of this part

Content:

The results part fit every requirement.

Language:

Similar general considerations as for the Introduction part stand here (Cf. above):

- **Lines 215, 216, 218, 223, 235, 241-244, 268** All the prompts should be written in italics as in a foreign language (Dutch).

  Thank you for your suggestion. We put the queries in italics and added the translation. However, for lines 223, 235, and 268 the term is the same as the translation, therefore we feel that it does not need to be translated. For lines 241-244 we made the change for those instances where the translation deviated from the original. We did write the queries in italics now as suggested.

- **Line 210** In English writing, the separator for decimal numbers is a dot “.” and not a comma “,” (9.7%).

Thank you for noticing this, indeed the comma should have been a dot here, we have now changed this (also in the other instances where we made this mistake).

- **Line 284** Figure 3 The legend “not in AGNES” is not clear to the reader.

  Thank you, to clarify this we have changed the “not in AGNES” to “Found previously but not in AGNES”.

Minor issue:

- **Lines 210 – 211** The numbers should be kept, and the per cent written under parenthesis 439 relevant hits (9.7%; Table 4)

  Thank you for noticing this, indeed the comma should have been a dot here, we have now changed this.

- **Lines 220, 238, 272** Table 4, Table 5 Table 6 Add a column for the percentages.

  Thank you for your suggestion, we have added the columns with percentages.

- **Lines 231-234, 248-250** Figure 1, Figure 2 Add, if possible, the percentage of each node.

  Thank you for your suggestion, however we feel that adding percentages here would clutter the graph, making it illegible. Following the above comment however percentages can now be found in the tables for interested readers.

- **Line 259** Add the percentage for the 89 sites under parentheses.

  Thank you for your suggestion, as suggested elsewhere we now added percentages to table 6. However, we feel that adding percentages for the 89 sites mentioned here will create confusion as this sentence mainly refers to the thirteen sites mentioned in the first part of the sentence. Therefore, we feel that if we add a percentage for the 89 sites (which are 56.3% of the total number of sites) the reader might wonder why a percentage is added here, but not for the thirteen sites. If we however add a percentage for the 13 sites it will also be unclear to the reader whether the percentage is a percentage of the total number of sites (n=158) or a percentage of the previously mentioned 89 sites. To prevent confusion, and because it would add unnecessary repetition (as the percentages are now mentioned in table 6), we decided not to add a percentage here.

- **Line 284** Figure 3 Why choose a palaeogeographical map when the environment is not discussed in the discussion part? The coordinates and projection systems are required on the left map.

  Thank you for this comment. We agree that the palaeogeographical map is not strictly necessary. To accommodate this comment we changed it to a more neutral map. We also included the coordinates and projection system in the new maps (for both figure 3 and 4, and for the newly added figure 1).

Discussion and conclusion:

Major issue:

- *Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? [ ] Yes, [x] No (please explain), [ ] I don't know.* I would have liked a more detailed discussion on the implication of such "by-catch" techniques for further research in grey literature (**l.379-395**). Only a few lines describe future possibilities, while there are many, and can also overcome strong bias, such as the absence of specialists in some excavations, which led to underestimating or misinterpreting some scattered findings. With this approach, this bias could be counterbalanced.

  Thank you for your comment. We do address this to a limited extent, also in our conclusion: "This type of 'by-catch' cannot be effectively found through other means, therefore it is recommended that AGNES is used systematically in tandem with established search methods." The reason that we kept potential future uses vague is because the relevance is case dependent (something we notice when we compare the present study to the previous study by Brandsen and Lippok). Hence we do address this issue mainly relating specifically to our own case study. For example in the following section:

  "A core strength of AGNES is that it allows us to find 'by-catch' in archaeological reports. Many of the previously unknown Vlaardingen Culture sites found in AGNES can be considered to be 'by-catch'. This is for example the case with the site Nijmegen Park Waaijenstein. The excavation focussed on a Roman period settlement in the area. The metadata of the report in DANS only mention the Roman period settlement. The Vlaardingen Culture remains at the site consist of three ceramic sherds and a few flint artefacts (Daniël 2018). The site Bergharen de Weem presents a similar case, the report titled “On the edge of a mediaeval settlement”[1] is focussed on mediaeval finds. The metadata in DANS mention Neolithic remains but those are not specified. As such they would not be found with the queries such as “Vlaardingen Cultuure” (Vlaardingen Culture). Only in the full text of the report is it specified that these remains consist of flint and ceramics from the Vlaardingen Culture (Diepeveen & Van Enckevort 2009). It is not surprising that many of these by-catch finds are located in the region of Nijmegen. Nijmegen is the heart of the Roman Netherlands, it is the oldest city in the country, and a major centre during the mediaeval period. Archaeological excavations frequently yielded large quantities of finds from these periods. It is perhaps unsurprising that a handful of Vlaardingen Culture sherds or flint artefacts on these excavations do not end up in the metadata of these reports. This is also no longer problematic as we were now able to retrieve this kind of information through AGNES.”

In our view the suggestion, mentioned by the reviewer, that research biases, in the absence of specialists at a site, might be overcome by these full text searches is plausible. But it does not follow from our study. We did not observe that the lack of specialists was a reason why by-catch was not mentioned in the metadata. As described in the section quoted above we mainly observed that by-catch was not mentioned in the metadata because the VLC remains were quantitatively limited compared to remains from other periods. This is illustrated well by the Bergharen de Weem site mentioned previously. For this site the flint was described by Elly Heirbaut, a lithics specialist. Therefore, we feel that the lack of specialists was not the reason why the Vlaardingen Culture remains were not mentioned in the metadata. Rather, because of the extensive medieval remains the metadata was geared to describing those remains. To conclude, we feel that speculating about such potential biases does not follow logically from our own observations, and we feel that this would fall outside of the scope of the present paper.

- *Are the conclusions adequately supported by the results (without overstating the implications of the findings)? [x] Yes, [ ] No (please explain), [ ] I don’t know*

Structure:

The general structure of the chapter is clear except for **lines 309-313**, which would fit better in the introduction as they define the different cultures.

Thank you for your suggestion, we agree that the Stein group should already be introduced in the

[1] “Aan de rand van een middeleeuwse nederzetting” (Diepeveen & Van Enckevort 2009).

introduction, rather than here. We now moved these sentences (and slightly altered them to fit the introduction) and added the following sentence to the introduction "Culturally the group is closely associated to the Stein group which is mostly located in the Limburg area (see figure 1)." We also added figure 1 to highlight the geographical range of the Stein group.

We now slightly changed the first sentence of paragraph 4.1. to make it fit better:

"Regarding the cultural attributions of Vlaardingen Culture sites we decided to adhere to the conclusions presented by the excavators."

Lines 357 – 361 would also refer to the results and not the discussion part.

Thank you for your comment. While these sentences indeed refer back to the topic discussed in the result section we felt they were necessary here as they provide the setup for the following discussion in which we explained/discussed these observed differences. We feel that moving these to the results section would not be helpful as the start of the paragraph would be too abrupt.

In the conclusion, the paragraph from **lines 444 – 449** does not entirely fit into the conclusion and would need to be either rewritten or changed into the discussion part.

Thank you for your comment, however, we do believe this section needs to be in the conclusion because it summarizes an important finding, namely that different terms are applied depending on the geographical location, irrespective of the archaeology at those sites. These sentences thus summarize the conclusion of paragraph 4.1.

Content:

There is one general lack of development of the discussion. As a non-specialist of the VLC, I cannot provide information on whether the newly founded site would improve our knowledge of this culture. However, the interpretation and possible uses of "by-catch" are limited to a few lines (**l. 390 -395**), while its possibilities extend to many areas and timelines and could help fix bias from the survey, in particular when specialists are missing.

We do end our conclusion with the following suggestion regarding future studies:

"This type of 'by-catch' cannot be effectively found through other means, therefore it is recommended that AGNES is used systematically in tandem with established search methods."

But as mentioned before, we feel (and observed both in this study and the previous study by Brandsen and Lippok) that the added benefits of this by-catch are case dependent. The suggestion that lacking specialists might create biases does not follow from our data. While it is plausible that this can be a problem in other case studies we feel that such a hypothetical suggestion does not fit in this article as we have opted for a data driven approach in which we focussed on addressing issues which followed from the presented case study.

Another comment on the F1 score (**l.397 – 398**) is whether it would be possible to recall the already identified sites (Found previously and in AGNES = 39).

We have added a sentence describing the recall for just the known sites.

Language:

Similar general considerations as for the Introduction part stand here (Cf. above):

- **Lines 398, 401** In English writing the separator for decimal numbers is a dot "." and not a comma ",".

  Thank you for the comment, we have changed this.

- **Lines 426-428** An unclear sentence and a repetition of the word "aimed".

  Thank you for your suggestion we changed "which was aimed at finding" in this sentence to 'which attempted to find'.

- **Line 428** The word "aimed" is used once again.

  As noted above we changed the second use of the word 'aimed'.

- **Lines 460 – 461** Redundant sentence.

  Thank you, we removed the sentence.

Minor issue:

- **Line 329** Figure 4The same comments are made as for figure 3.

  Thank you for your comment. We agree and removed the palaeogeographic map as suggested before for figure 3.

- **Line 418** A reference or hyperlink would be needed for *PyMUPDF* software.

  We have added a link

- **Line 431** Thirty and its percentage.

  Thank you we have added this.

- **Line 459** The number before the percentage.

  Thank you we have added this.

Literature:

The choice of literature seems quite reasonable overall. I would only suggest recent publications on the uses of LLMs in archaeology (Agapiou and Lysandrou, 2023; Cobb, 2023; Lapp and Lapp, 2024), and more especially the book of Gonzalez-Perez et al. (2023) Discourse and Argumentation in Archaeology: Conceptual and Computational Approaches with several chapters on NLP or text extraction.

Thank you for the useful suggestions, we have added these citations to the paper.

Conclusion:

In conclusion, as such, I cannot recommend this paper for publishing. It needs major revisions. The problems reside in the confusion of the large language model, the lack of context on the Vlaardingen Culture, and the methodology workflow needing to be more transparent.