Author Name Disambiguation using Large Language Models

Contributions to a system for open reproducible publication research

Jelle van Lieshout[1]

Supervisor(s): Diomidis Spinellis[1]**, Georgios Gousios**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 28, 2024

Name of the student: Jelle van Lieshout
Final project course: CSE3000 Research Project
Thesis committee: Diomidis Spinellis, Georgios Gousios, Koen Langendoen

Abstract

Author name disambiguation, otherwise described as (publication) record linking, is a problem that has had considerable research dedicated to its solving. Author attributions, calculating research metrics and conducting literature reviews are amongst processes that experience increased difficulty due to ambiguous author names. In this study, a novel approach is presented to disambiguate authors related to scientific publications, using Large Language Models (LLMs) in combination with the Alexandria3k software package. LLMs have shown great potential in processing, analysing and drawing conclusions when presented with human-readable data. The approach presented in this study supplies a LLM with known attributes of publication records and authors, such as names, affiliations and co-authors, to determine whether records written by authors with ambiguous names can be linked to the same real-world person. Using Alexandria3k, a dataset of authors and publications with confirmed identities is created to test and validate the approach. Finally, the approach is measured against state-of-the-art methods to disambiguate author names and different configurations are presented and discussed.

1 Introduction

Systematically combining results from different studies, otherwise known as research synthesis, allows one to provide a more comprehensive understanding of a particular question or area of research. Researchers and scientists can, using this methodology, build on top of existing knowledge, drawing conclusions from pre-existing research and publications. One way to retrace the references mentioned in a publication, is to use online data sources such as bibliographic and article databases. However, performing systematic studies on published literature through the available online systems can be problematic [13]. Experiments and related data can be difficult to reproduce and are often not transparent.

To address this and other issues, Alexandria3k has been created. Alexandria3k is a Python software package and an associated command-line tool that can populate embedded relational databases with slices from the complete set of several open publication metadata sets for reproducible processing through sophisticated and performant queries [13]. A system designed to process data is hugely dependent on the quality of its input data. In the case of Alexandria3k, one occurring issue is that author names are not exclusively unambiguous. While providing a solid foundation, it is desirable that Alexandria3k's functionality is therefore further researched, specifically within the scope of name disambiguation.

During recent years, usage of open-source and commercially available LLM's has greatly increased. Examples of such models are GPT-3.5, Phi-2 and Llama-2 [10]. Large language models have shown great potential recognizing patterns and deriving and understanding of textual content. In this study, we will conduct experiments to establish whether this potential can be harnessed and used to solve the problem of ambiguous author names.

An approach is outlined to effectively disambiguate author names using LLMs, within the infrastructure of Alexandria3k. While the latter provides us with ample publication data as well as informative attributes to those publications, this information can effectively be processed and authors can be compared using the power of LLMs. In the spirit of Alexandria3k, experiments will be run using LLMs available to and usable on consumer hardware.

2 Background and Related Work

The two most outspoken topics appearing in this study, large language models and author name disambiguation, have both been thoroughly researched. Large language models are a more recent finding in comparison to the problem of author name disambiguation, but nevertheless have had considerable research dedicated to its applications and inner workings. Author name disambiguation is in itself a problem that is essentially as ancient as the first documented research publications. Different approaches have been researched and proposed, often influenced by novel findings in other fields. To explore where our novel approach lies within the ecosystem of existing methods, we will use a taxonomy specifically designed for author name disambiguation methods, as described in in an article by Ferreira et al [4], shown in figure 1. The described taxonomy provides a rough idea and overview of what other type of methods have previously been researched.

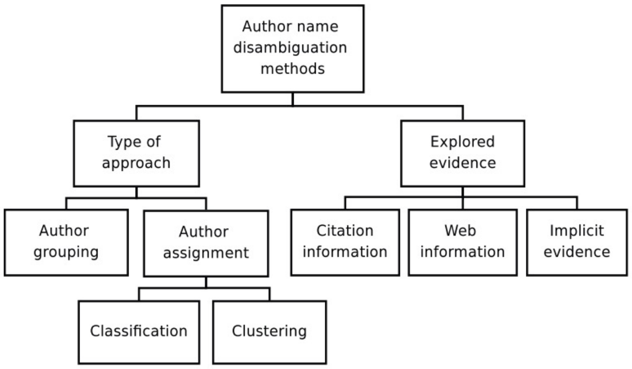
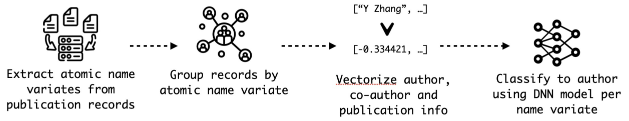


Figure 1: Author name disambiguation methods taxonomy as described by Ferreira et al [4]

The large language model based method described in this study presents an interesting case when held against the taxonomy model. First, if we take a look at the "type of approach", we can conclude that the large language model is a type of "author grouping". In short, because the approach considers two publication records and either links them onto the same real-world author, or not. The large language model approach also contradicts both definitions within the "author assignment" category as described in [4]. If we focus our attention to the right branch of the taxonomy, we can conclude that when it comes to explored evidence, the large language model approach intends to use all three described categories. ”Citation Information”, as this is what the model is prompted with, ”Web Information” because this is potentially included in the dataset used to train the model, and ”Implicit evidence” due to large language model’s ability to detect patterns and draw information from context.

Author name disambiguation is a problem to which many solutions have been proposed through studies [3][9][4]. Solutions range from, but are not limited to, clustering techniques, supervised learning, unsupervised learning and rule-based approaches. The most recent study conducted in 2023, describes an approach using a neural network called *WhoIs*, applied to the DBLP repository dataset. It distinguishes authors based on relationships with co-authors, area of research and titles of publications. WhoIs was created and described by Boukhers and Asundi and will be further referred to as such in this study[3]. In the experiments described in the study, the Boukhers and Asundi approach outperformed state-of-the-art approaches on the DBLP dataset[3]. To our knowledge, this is the most effective approach to disambiguate author names at this time. Returning to the previously discussed taxonomy, the Boukhers and Asundi approach could be described as a classification method using citation information. Due to the shown advantageous effectiveness of the Boukhers and Asundi approach, we will use this method as our benchmark to compare the novel approach against. Figure 2 shows a diagram outlining the process the Boukhers and Asundi approach follows.



[“Y Zhang”, …]
[-0.334421, …]
Extract atomic name variates from publication records
Group records by atomic name variate
Vectorize author, co-author and publication info
Classify to author using DNN model per name variate

Figure 2: Boukhers and Asundi approach represented in a flow-chart

LLMs and their advancements have not gone unnoticed in recent years. The most recent advancements have provided us with LLMs that are able to complete several natural language processing related tasks, such as answering questions or translating language. Studies have been conducted related to documenting LLM development and capabilities [17], but not yet for the specific purpose of author name disambiguation. As a solution to disambiguate author names, we will run a series of experiments to establish whether the use of LLMs is beneficial in this process. LLMs have shown their strength in a number of applications, having the ability to extract information and discover patterns in a similar fashion to what a human can determine.

3 Methodology

The most recent and to our knowledge most effective approach to disambiguate author names at this time, is a deep neural network based system described by Boukhers and Asundi [3]. In short, the system groups publication records by atomic name variate (e.g. ’Paul Smith’, ’Patrick Smith’ and ’P. Smith’ all get grouped under the atomic name variate ’P. Smith’), and subsequently trains a deep neural network using embeddings of previously identified author names, co-author names, publication titles and journal names. A ’new’ publication record belonging to the same atomic name variate group, can then be classified to a real-world author using the deep neural network.

While implemented, both approaches will be tested and compared using comparable datasets, referring to one single atomic name variate. Based on the results, a precision (positive predictive value), recall (sensitivity) and F-score (combined accuracy measure) will be calculated for every conducted experiment. Based on the calculated scores, we can compare and establish whether the LLM approach is indeed more effective and accurate than the current state-of-the-art approach.

To test the accuracy of any method, one needs a dataset as well as a ground truth to measure whether the method has come to a correct or an incorrect conclusion. For this purpose we will make use of the CrossRef [6] dataset. CrossRef is a non-profit organization maintaining a registry containing more than 134 million publication records. The previously described Alexandria3k system will allow us to easily handle and operate on this data. A subset of this data contains authors that have been identified using an Orcid ID, allowing use as a ground truth for the experiments conducted.

In an attempt to test large language models’ abilities in general, we will conduct experiments using three different large language models. Each individual model will have details about the amount of parameters it has been trained on. In general, one can assume that when the same model is trained on a larger amount of parameters, it can perform better on certain tasks. However, performance between different models can deviate while being trained on an equal amount of parameters [2]. LLMs have a large set of configuration parameters that can be adjusted. Experiments are also conducted to achieve an optimal configuration. There are two parameters of which we are trying to ascertain their relevance. Firstly, LLM temperature, which affects the randomness or unpredictability (often intepreted as creativity) of the model [8]. Secondly, we alter the different informative aspects related to the publications the model can consider. It should be noted that this study attempts to reason about LLMs ability to disambiguate author names in a general sense, and there are limitations in terms of the LLM size that we are able to experiment on within reasonable time spent.

Llama2 13B [15], *Mistral 7B* [7] and *Zephyr 7B* [16] are all dialogue-based large language models. Llama2 was developed as part of a set of models ranging from 7 to 70 billion parameters by Meta, and is trained on 13 billion parameters. Mistral is a model trained on 7 billion parameters, and Zephyr is a fine-tuned version of Mistral 7B using Direct Preference Optimization, also trained on 7 billion parameters, allegedly aligning the models output with the users intent to a greater extent [11]. All three models have similar reasoning and information extraction capabilities in a general sense, and can be run on consumer hardware.

When studying effectiveness and accuracy of the approach proposed in this study, we will use the system described by Boukhers and Asundi as a baseline to compare against. Both the new approach as well as the Boukhers and Asundi approach are implemented within Alexandria3k, and trained, validated and tested using publication records imported from CrossRef [6]. The publication records used, have been filtered and are only used in case their corresponding authors have already been identified using an ORCID iD [5], to use as a ground truth. Due to the large size of the CrossRef dataset as a whole, we will run experiments on a 10% sample of randomly selected publication entries, corresponding to authors whose ORCID iD has been previously specified. Alexandria3k supports this operation by default, using the

```
--sample 'random.random()' < 0.1'
```

command flag [12]. The sample database is established and popuulated using the following command:

```
    a3k -d populate crossref_sample.db crossref
"April 2022 Public Data File from Crossref/"
--row-selection "work_authors.orcid is not null"
--sample 'random.random() < 0.1'
```

The experiments are conducted within the Alexandria3k system. Alexandria3k gives us practical access to publication record datasets in relational form, and provides an infrastructure that allows the experiments to easily be reproduced. Additional features have been developed for the purpose of this study to train, validate and test both the Boukhers and Asundi approach [3] as well as the novel approach described in this study. The consequential calculation of the precision, recall and F-score are not features developed within Alexandria3k. With regards to our measurement statistics for the novel approach, precision, recall and F-score, we use the following definitions for true and false positives and negatives.

True Positives (TP) - publication records matched to another publication record with the same real-world author according to our ground truth.

False Positives (FP) - publication records matched to another publication record with a different real-world author according to our ground truth.

True Negatives (TN) - publication records not matched to another publication record with a different real-world author according to our ground truth.

False Negatives (FN) - publication records not matched to another publication record with the same real-world author according to our ground truth.

Using the definitions of true and false positives and negatives, precision, recall and F-score are consequently calculated as follows.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{FP + FN}$$

$$Fscore = \frac{2(Precision * Recall)}{(Precision + Recall)}$$

A perfect approach would score 1.0 on each of the three above mentioned indicators.

3.1 Approach

As described, we will use several LLMs to disambiguate author names. Different from previously researched classification and clustering approaches, we will use the LLMs ability to recognize patterns and understand context to ascertain whether two publication records with identical atomic name variates, could be written by the same real-world author.

For our purpose, we use a pre-trained transformer model, that is trained on a vast and diverse dataset consisting of text from all types of sources. The model is then provided with a *system prompt*, containing its operation guidelines (a Zephyr-based example is shown in Appendix A). These guidelines include the task at hand: when provided with two publication records, determine whether these two publication record could be linked to the same real-world author. It also includes instructions on how to receive input, as well as in what specific format to provide answers. The system, when provided with two publication records, operates according to the following steps. In figure 3, the approach is shown visually.

1. **Preparation and pre-processing.** Extract relevant information from publication records: author first and last name, co-authors first and last name, journal name, affiliated organization name, publication title and article subjects.

2. **Prompting and interpreting output.** Present combinations of two publication records to the model and consequently parsing and mapping its output to a match or non-match.

3. **Log and save.** After receiving a positive or negative result from the model, log and save the result.

4. **Repeat.** Repeat previous steps until there are no publication records combinations corresponding to authors left, repeat the first three steps.

5. **Validate and calculate metrics.** Validate matches using the ground truth [5], and calculate the desired metrics.

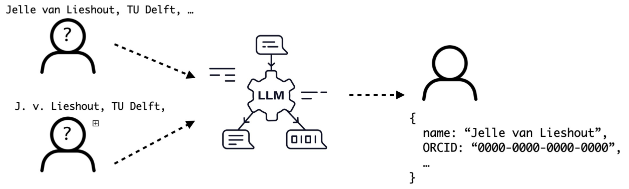

Jelle van Lieshout, TU Delft, …
J. v. Lieshout, TU Delft,
LLM
{
name: “Jelle van Lieshout”,
ORCID: “0000-0000-0000-0000”,
…
}

Figure 3: Diagram of LLM Approach

While the Boukhers and Asundi approach [3] operates on an entire set of publication records referring to a single atomic name variate, our LLM based approach makes use of the fact that in many cases, author first names are known. We separate the publication records within an atomic name variate set by their author first name, and present all possible combinations of publications within that set to the model. Information included in the presented publication object is shown in listing 1.

```
{
    "title": "Author Name Disambiguation
    ↪  using Large Language Models"
    "published_year" : "2024",
    "container_title" : "Electronic
    ↪  Repository Delft University of
    ↪  Technology",
    "short_container_title" : "TU Delft
    ↪  Repository",
    "publisher": "IEEE",
    "co_authors" : ["John Doe", "Jane
    ↪  Doe"],
    "subjects": ["Computer Science"],
    "affiliated_organizations": ["TU
    ↪  Delft"]
}
```

Listing 1: JSON example

3.2 Experimental Setup and Results

To compare both the novel LLM based method and the existing state-of-the-art deep neural network approach, as described in the first paragraph of this section, will use a 10% sample dataset consisting of random publication records from the CrossRef dataset, with previously identified authors. This corresponds to roughly 733.000 publication records, 1.112.100 authors and 1.017.143 unique author names. From this set, we can then use author atomic name variates with many occurences to compare the methods. An example of such an atomic name variate is ”Y Zhang”. Table 1 shows the results of the application of the Boukhers and Asundi approach on this atomic name variate.

Name variate	Publications	Precision	Recall	F-Score
”Y Zhang”	416	0.857	0.857	0.861

Table 1: Results using the Boukhers and Asundi approach on atomic name variate ”Y Zhang”

The ”Y Zhang” atomic name variate has the most corresponding publication records within the CrossRef dataset, we will therefore use it as a representative sample in our experiments. In our sample dataset, the ”Y Zhang” name variate contains 2770 publications. As shown in table 1, only 416 of those publications are used to test the Boukhers and Asundi approach. The leftover publications had been used before this to train and validate the neural network. For the novel LLM approach, 1161 out of 2770 publications records fulfill all the requirements in terms of necessary informative attributes to accurately run the model.

Pre-processing of publication records

An important thing to note, in terms of experiment setup, is that there are limitations as to what models can reasonably be run on consumer hardware. In this specific environment, experiments will be running on a MacBook Pro (2024) with 18GB of random access memory. An experiment where 1000 publication combination comparisons are made, takes approximately 2 hours. One comparison using e.g. Zephyr on the machinery the experiments where conducted on, takes roughly 3 to 5 seconds. The developed LLM approach therefore executes a number of pre-processing steps to assure only necessary comparison are done.

1. **First name equalization.** First names are converted to lowercase strings, and all spaces and dashes are removed, such that e.g. ”Yi-Quan” and ”Yi Quan” both become ”yiquan”.

2. **Comparison selection.** Only publication records with identical first name entries are compared. This removes the opportunity to match some publication records that are actual matches, but is a trade-off made to reduce computation time.

3. **Size reduction.** The LLMs that have been used, have a limited context size of roughly 32000 characters that can be processed at once. To assure that the proposed publication combinations do not exceed this context size, some attributes strings are shortened.

Finding the right configuration

To test for real-world author correspondence when presented with an author name with n related publication records, $\binom{n}{2}$ comparisons are conducted. Because comparisons are costly in terms of time and computational power, we attempt to find a usable configuration using a small yet challenging benchmark sample. This sample consists of 3 different publication sets corresponding to 3 arbitrarily chosen author names from 3 different author segments. One segment contains publications with 80% to 100% unique authors, one segment has a unique-author-to-publication ratio of 40% to 60% and one segment contains publications all written by the same real-world author. Author segments considered are limited to 20 publications, given the large increase in comparisons needed otherwise. We will use the smaller benchmark set to run experiments and determine an effective LLM configuration to be used on larger datasets. The arbitrarily chosen author names for the configuration experiments and their corresponding metrics are shown in table 2 . To determine what configuration suits our purpose best, a number of experiments was conducted on this smaller benchmark sample of publications.

Name	Unique Authors	Publications	Comparisons
Yi-Quan Zhang	1	12	66
Yingjie Zhang	6	13	78
Yifan Zhang	17	19	171

Table 2: Benchmark publication sets selected for configuration experiments

In figure 4 the averaged metric results are visible upon altering the LLM temperature parameter, when prompted with publication records with no informative attributes omitted. From the figure, one can observe that Llama2 and Zephyr are quite affected by the temperature parameter, while Mistral obtains quite consistent results. It should also be known that, the higher the temperature parameter is set, the more disobedient (i.e. not following system prompt instructions) Llama2 becomes when providing output. This makes parsing and concluding a match / non-match significantly more difficult.

In terms of temperature, we propose two strong candidates for the larger dataset: *Mistral with a 0.0 to 0.6 temperature setting* and *Zephyr with a 0.6 temperature setting*. Both Mistral and Zephyr clearly outperform Llama2 with the exception of recall in the case of temperature 0.8. Llama2 will, regardless of lesser performance, still be taken into consideration in further experiments.

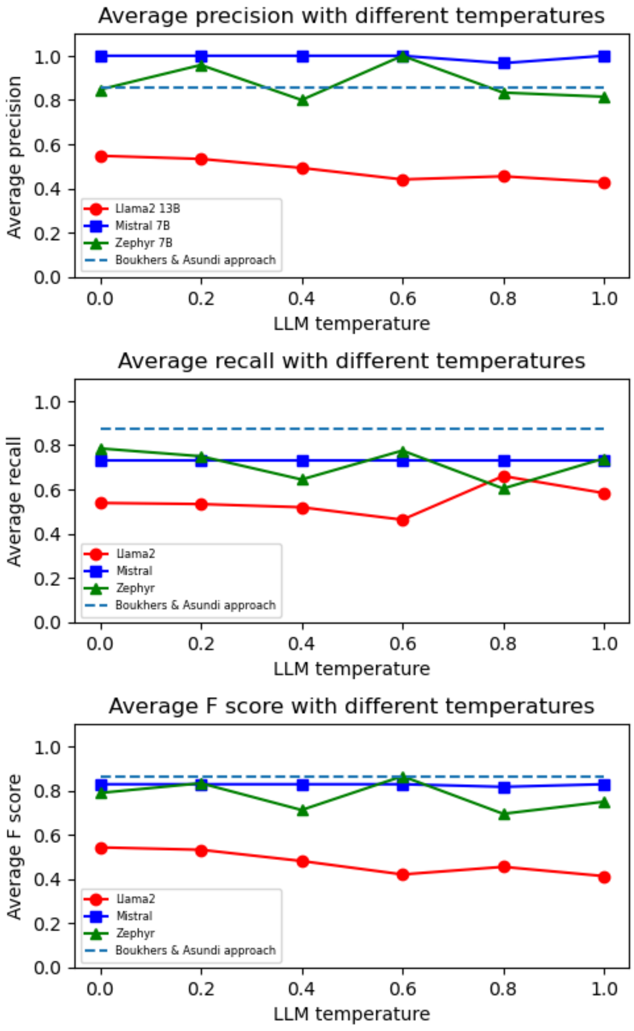


Figure 4: Average precision, recall and F-score with different LLM temperatures

In figure 5 the averaged metric results are visible when omitting certain informative attributes when presenting publication records to the LLMs. Experiments are conducted while information is omitted regarding *affiliated organizations, subjects, co-authors and publication title*. It is important to note that the described omitting of information is twofold: the actual informative attributes are removed from the publication record object, and the relevant system prompt provided to the LLM is also adjusted such that no instructions remain to predict based on the omitted attributes.

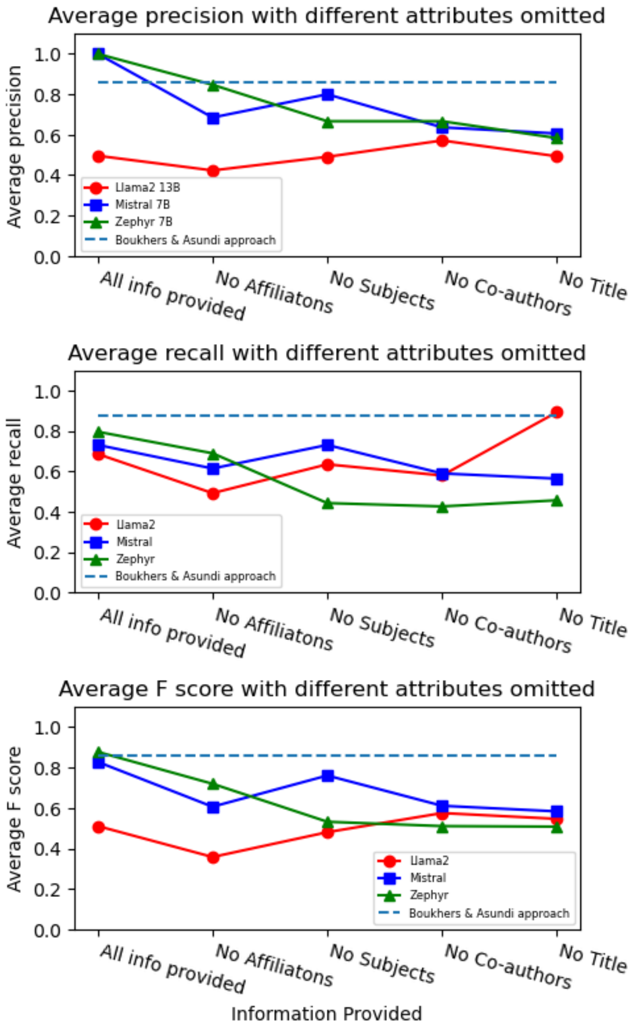


Figure 5: Average precision, recall and F-score with different informative attributes omitted

For both Zephyr and Mistral, we can conclude that the approach benefits from having access to as much relevant attributed information as possible, when comparing publication records. The results produced by Llama2, indicate that performance is roughly equal when providing all information versus leaving out the publication title. We can therefore include all publication information in futher experiments on larger datasets.

Results

With the results of the previous section in mind, three suitable candidates for the purpose of author name disambiguation have been selected: *Mistral with temp=0.0*, *Llama2 with* *temp=0.0* and *Zephyr with temp=0.6*. All models are tested against the baseline approach as described by Boukhers and Asundi [3]. To compare both approaches, publications corresponding to the name variate ”Y Zhang” are tested. It is important to note that not all publication records include the necessary information and attributes to be classified using the LLM approach succesfully, and are therefore not considered in this comparison. The LLM approach performance can also be negatively affected by the data pre-processing, as some (potentially matching) comparisons are discarded in pre-processing to save time and computational energy.

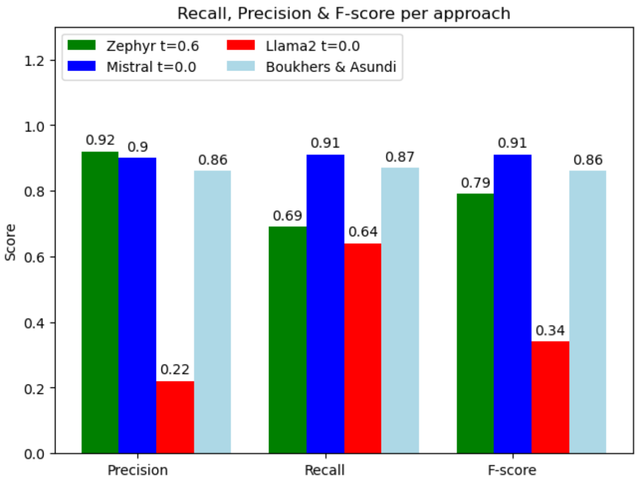


Figure 6: Precision, recall and F-score with different approaches

The graph shows the results of the two LLM approaches compared to the existing Boukhers and Asundi approach. Some of the missed matches by the LLM approach can be caused by the pre-processing steps. The results clearly show that overall, *Mistral with t=0.0* is the most effective approach from the three considered candidates, outperforming the Boukhers and Asundi approach on all three metrics.

4 Responsible Research

Several considerations should be noted for this study in the context of responsible research. As an established measure, we will use the Netherlands Code of Conduct for Research Integrity [1], and discuss the sections relevant to this study. Further in this section we will also elaborate on general considerations regarding the use of large language model and its consequences.

In conducting and this writing this study, the Netherlands Code of Conduct for Research for Research Integrity was followed where necessary and possible. Principles originating from the Code of Conduct that are, according to our knowledge, especially relevant when handling large amounts of data are: Honesty, Scrupulousness, Transparency and Responsibility. Honesty, Scrupulousness and Transparency have been assured by thoroughly testing and documenting the methods used in this study. The research methods used have also been discussed and assessed by peers within the same research group.

Responsibility, partially defined as ”acknowledging the fact that a researcher does not operate in isolation and hence taking into consideration - within reasonable limits - the legitimate interests of human and animal test subjects, as well as those of commissioning parties, funding bodies and the environment” [1]. To discuss this principle we focus mainly the human authors affiliated to the publication. In general, the data used in this study is publicly available and has also been published by the authors behind the publications with public availability in mind. For these publications, disambiguating their respective authors does not conflict with their intentions and should therefore cause no harm. There is an additional group of authors to be considered. Part of the reason author name disambiguation is necessary, is due to authors publishing books or articles under different names, with the intention to remain anonymous. Relating these publications to their respective real-world authors, causes this anonymity to vanish, and could therefore be harmful for the respective author. The approach described in this study, however, is largely ineffective on publications where no additional data regarding co-authors and affiliated organizations is provided. It should therefore not interfere with authors publishing with the intention to remain anonymous.

5 Discussion

The outcomes of the several experiments conducted, show that LLMs are a suitable candidate to solve the problem of author name disambiguation. It can, additionally, serve as a useful addition to an existing approach. For example, when working with a neural network approach, entries that have been classified with a confidence score below a certain threshold can be post-processed and classified using the LLM approach. Existing neural network approaches also tend to be ineffective when attempting to classify author entries with no previous work [3]. The LLM approach is not affected by an author having no previous data, and will most likely output a (correct) non-match when comparing with authors that are well represented in a dataset. On top of that, our experiments have shown that even on its own, LLMs are very suitable tools to disambiguate author names, surpassing the state-of-the-art approach in terms of precision, recall and F-score.

5.1 Implementation

Part of the goal of this study, alongside researching the capabilities of LLMs when it comes to disambiguating author names, is expanding the Alexandria3k package. While the approach described in this study can definitely be a valuable addition to the system, there are some caveats. One of the strengths of Alexandria3k, is that any user is able to download and run it on their personal machine. The procedures involving running LLMs locally on ones personal computer, have certain hardware requirements that can simply not (yet) be fulfilled by any standard laptop or PC, and when these requirements are fulfilled, disambiguating a single atomic name variate takes roughly 2 hours per 1000 comparisons. This causes limitations in implementability, though still within the realm of hardware available to consumers. A positive aspect, however, is that the LLMs used are pre-trained. It does therefore not require a large dataset to train e.g. a neural network.

The minimum dataset size to (effectively) apply the LLM approach is merely two publication records. This can be a benefit in the context of Alexandria3k as well, given that the approach can be used regardless of imported dataset size. Another solution to this issue could be to create a dataset of disambiguated publication records, and share this dataset online as a reference and additional datasource that can be imported using Alexandria3k.

5.2 Limitations

Similar to other approaches, the LLM approach performs poorly when handling cases where e.g. two publication records have ambiguous co-authors. However, these occurrences are rare.

Another limitation of the LLM approach is that, as shown in the experiments, its performance reduces significantly when certain informative attributes are missing. It should be noted that more additional informative attributes could be added to the LLM prompts, to reduce the effect of incidental missing attributes.

As discussed in earlier sections, every LLM comparison is costly in terms of time. We therefore do not necessarily have the luxury of comparing every possible combination of publication records. This limitation can partially be countered by pre-processing and pre-selecting likely matches as comparison candidates.

6 Conclusions and Future Work

Based on the results and considerations described in this study, we conclude that large language models are a very viable option to solve the problem of author name disambiguation. The LLM approach outperforms a state-of-the-art neural network based approach described by Boukhers and Asundi [3], and has very high precision, recall and F-score in general sense. Specifically *Mistral 7B* shows extremely promising results, outperforming the Boukhers and Asundi approach on all metrics, while being a relatively small LLM. Further research is recommended to evaluate whether using a larger (or smaller) LLM can increase precision, recall and F-score even more.

In the previous section, some limitations in terms of performance of the approach are discussed, as well as some practical advantages over other approaches. We conclude that LLMs present a practical and unique means to disambiguate author names, and a valuable addition to the already large ecosystem of solutions. The LLM approach excels specifically in terms of precision, recall and F-score, as well as the ability to operate with no threshold in terms of dataset size.

6.1 Future work

In this study the LLM approach was outlined, and has shown promising results. In this section some potential ways to improve the approach are discussed.

Pre-processing and pre-selection of candidates

Due to the costly nature of a publication comparison, the approach can benefit greatly from more effective pre-processing and selection of match-candidates. A rough implementation to do so is outlined in this study (essentially equalizing author first names and only comparing combinations with equal first name entries), but a lot of potential matches are still not covered using this approach (e.g. publication records where the first name is incorrect or only the first letter of the first name is provided). On top of that, the attributes gathered to be added to the prompt could be extended. Plenty of information regarding the publications is available within Alexandria3k, and it could very well be that more informative attributes increase the chance of an accurate author match.

LLM configuration parameters

In this study we have conducted experiments with different temperature parameters. It should be noted that temperature is not the only configurable parameter for LLMs, and that only one *system prompt* was used to run the experiments. The approach could potentially benefit from other parameter adjustments or a different *system prompt*.

LLMs trained specifically on scientific publication datasets

For the purpose of this study, we have used generic, readily available models that are operable on consumer hardware. It should be noted that models also exist, that are trained specifically to harbor reasoning and knowledge surrounding scientific publications. Examples are BLOOM, a 176B parameter model [19], and Galactica [14], both large language models specifically trained for the purpose of science. These models could inherently know more about the publications proposed to them, and match more accurately.

Chain of thought prompting

Except for the use of a *system prompt*, no advanced techniques have been used to improve the prompting sequences. One technique that has shown to improve result accuracy in other applications, is so-called chain of thought prompting. It essentially revolves around prompting step by step, while using examples, to help the LLM understand the desired output [18]. Techniques as such could potentially improve the accuracy of the LLM approach.

A LLM Model Configuration

```
FROM zephyr:latest
TEMPLATE """{{- if .System }}
<|system|>
{{ .System }}
</s>
{{- end }}
<|user|>
{{ .Prompt }}
</s>
<|assistant|>
"""
SYSTEM """
As an author name disambiguation assistant,
your task is to determine if two given
publications could possibly be authored by
the same real-world author. Upon receiving
information about two publications, consider
the following criteria for your analysis:
```

```
1. Publication Topics: Check if there is an
overlap in the topics of the two publications.
Even a general overlap can indicate the same
authorship.

2. Co-Author Names: Look for any overlapping
full names of co-authors in both publications.
If there is at least one common co-author,
answer 'YES'. This is a strong indication of
the same author.

3. Fields of Study: If the publications are
related in their fields of study, this could
also suggest the same authorship.

4. Affiliated Organizations: Overlapping
affiliated organizations in both publications
can be a sign of the same real-world author.

Based on these criteria, provide a response
of 'YES' if you determine potential same
authorship, or 'NO' if otherwise. Please
restrict your answer to only 'YES' or 'NO'
to avoid confusion.
"""
PARAMETER temperature 0.6
PARAMETER stop "<|system|>"
PARAMETER stop "<|user|>"
PARAMETER stop "<|assistant|>"
PARAMETER stop "</s>"
```

References

[1] Netherlands Code of Conduct for Research Integrity — NWO.

[2] Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., et al. Scaling Language Models: Methods, Analysis & Insights from Training Gopher.

[3] Boukhers, Z., and Asundi, N. B. Deep author name disambiguation using DBLP data. *International Journal on Digital Libraries* (2023).

[4] Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. A Brief Survey of Automatic Methods for Author Name Disambiguation. *SIGMOD Record 41*, 2 (2021).

[5] Haak, L. L., Fenner, M., Paglione, L., Pentz, E., and Ratner, H. ORCID: A system to uniquely identify researchers. *Learned Publishing 25*, 4 (10 2012), 259–264.

[6] Hendricks, G., Tkaczyk, D., Lin, J., and Feeney, P. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies 1*, 1 (2 2020), 414–427.

[7] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., et al. Mistral 7B.

[8] Jiang, Z., Araki, J., Ding, H., and Neubig, G. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering.

[9] Mihaljević, H., and Santamaría, L. Disambiguation of author entities in ADS using supervised learning and graph theory methods. *Scientometrics 126*, 5 (5 2021), 3893–3917.

[10] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. A Comprehensive Overview of Large Language Models.

[11] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model.

[12] Spinellis, D. https://dspinellis.github.io/alexandria3k/, 2023.

[13] Spinellis, D. Open Reproducible Scientometric Research with Alexandria3k. *PLoS ONE 18*, 11 (11 2023), e0294946.

[14] Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science.

[15] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open Foundation and Fine-Tuned Chat Models.

[16] Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. Zephyr: Direct Distillation of LM Alignment.

[17] Usman Hadi, M., al tashi, q., Qureshi, R., Shah, A., muneer, a., Irfan, M., Zafar, A., Bilal Shaikh, M., Akhtar, N., Wu, J., Mirjalili, S., Al-Tashi, Q., and Muneer, A. A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. *Authorea Preprints* (10 2023).

[18] Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (2022), S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., pp. 24824–24837.

[19] Workshop, B., Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ili, S., Hesslow, D., Castagné, R., Sasha Luccioni, A., et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.