PROCEEDINGS OF SCIENCE

Easily accessible (small) LLMs in historical studies: opportunities, limitations, pitfalls

Marianne Groep-Foncke* **and Tine De Moor**

Rotterdam School of Management, Erasmus University Rotterdam,
Burgemeester Oudlaan 50, Rotterdam, Netherlands

E-mail: groep@rsm.nl, demoor@rsm.nl

Automation of certain parts of data collection and data processing in the field of Social Sciences and Humanities would enable researchers to skip some of the more painstaking tasks, leaving more time for the actual analysis and opening up the possibility to work with larger data sets. With the current and imminent generations of open-source Large Language Models (LLMs) and small Large Language Models (sLLMs) it seems already attainable for individual researchers to speed up onerous but necessary tasks such as qualitative data coding using personal computers, while keeping control of their datasets at all times. Using the qualitative coding process of the former Common Rules project of research group Social Enterprises & Institutions for Collective Action (SEICA, Erasmus University Rotterdam) as a case study, we explored whether open source, low-threshold (s)LLMs are already able to perform specific tasks such as qualitative data coding, and what are the limitations and pitfalls that current and future researchers in the field of Social Sciences and Humanities have to reckon with when applying such AI-driven aides.

International Symposium on Grids and Clouds (ISGC2025)
16 -21 March 2025
Academia Sinica Grid Computing Centre (ASGC), Taipei, Taiwan

*Speaker

 ISSN 1824-8039 . Published by SISSA Medialab.

PoS(ISGC2025)004

1. Introduction

Between 2011 and 2015, the research group Social Enterprises & Institutions for Collective Action (SEICA, Erasmus University Rotterdam) – at that time residing at Utrecht University as the research group Institutions for Collective Action – ran the Common Rules project (funded by the Dutch Research Council, grant number 236-50-005). Its aim was to gather insights on the development of efficient and effective regulations in past and present. The research group at large studies institutions for collective action, defined as "governance regimes of natural or man-made resources, shared and managed by groups of users according to collectively set agreements" [25]. The Common Rules project focused on a specific type of institutions for collective action – commons – in England, the Netherlands, Belgium, and Spain. Within these countries, the focus was on specific areas with a substantial concentration of pasture commons in the period before 1800. Commons existed in these areas from the Middle Ages through the nineteenth century and were set up to coordinate the individual use of collectively managed natural resources such as grassland, woodland, or water. As in all institutions for collective action, the members or shareholders of these commons collectively set the rules on, among others, access to membership, management of the collective resources, and use of both the resources and eventual services provided by the group. In several cases, these regulations were put down in writing, although it is probable that in most cases verbally agreed rules, norms, and values pre-existed the written rules. The research done during the Common Rules project and the subsequent MIDI project (Modelling Institutional Dynamics In historical commons, funded by the Riksbanken Jubileumsfond) provided novel insights into the resilience of institutions for collective action in general, the interplay between external changes and the institutional setup of such organizations, and the role of specific types rules and (the absence of) sanctions [9][10][13][21]. Farjam et al. demonstrated that commons, regardless of their legal background, followed a specific development path over time, showing a U-curve, whereby commons have a stronger regulation activity at the beginning and the end of their life span, with a higher incidence of administrative rules towards the end [13]. De Moor et al. highlighted that the lesser the regulatory activities that commons needed, the longer they survived as organisations [10]. Such insights on the longitudinal development on commons can contribute substantially to the ongoing debate on the management of collective resources, which received a very large boost when the Nobel prize in Economics was awarded to Elinor Ostrom, one of the founders of studies on commons who invested heavily in identifying the factors that contributed to the formation of resilient institutions for collective action.

To be able to discern such patterns among the rules issued by the selected commons, a lengthy and labour-intensive procedure was followed: the (hand-written) rule set of each common was transcribed, thereafter translated to modern English, each text block was meticulously deconstructed into "rule statements" and each element therein was coded according to a codebook developed by the participating researchers. Furthermore, the research group built a database with historical and geographic information of each common (e.g., years of existence, location, dimensions), details on the type of resources managed, governance arrangements, parties involved, rights and dues of shareholders, and specifics on the actual rules and sanctions (e.g., rule form, repetitiveness). In the end, the building of the database paid off, leading to the results mentioned above, but it needed the allocation of a huge amount of human capital. If we could use (s)LLMs or other machine-learning

techniques to do or at least support tasks like these in the future, this would free up time for our actual scientific analysis. Moreover, it would give us an opportunity to work with more and larger data sets. This suits the needs of the SEICA research group. Over the past decades, the research group's focus has expanded from European institutions for collective action in the Middle Ages and the early modern period to institutions for collective action from that period up till the present day, with projects such as UNICA (Building a UNified theory for the development and resilience of Institutions for Collective Action in the past millennium), SCENTISS (Boosting Social and Community-driven ENtrepreneurship for the Transition to an Inclusive and Sustainable Society), and the release of the CollectieveKracht.eu (CollectivePower.eu) knowledge exchange platform for present-day institutions for collective action. These and future projects – and other valuable work done in the field of Social Sciences and Humanities – would highly benefit from easily accessible, reliable tools that are able to advance painstaking but necessary tasks such as qualitative data coding. The objective of this paper is to replicate the qualitative coding process done in the past with the use of an sLLM, thereby exploring the possibilities and the limitations of applying ready-made sLLMs in historical research. The audience we aim at with this paper is twofold. Not only does it outline a use case of sLLMs in a specific scientific domain to computational scientists – highlighting particular characteristics to be reckoned within this domain – it also provides guidance for historical and other humanities researchers who consider using language models for qualitative coding processes. For other possible applications of sLLMs in Social Sciences and Humanities, see the excellent tutorial of Hussain et al. [19].

2. Related work

Several scholars have explored the capabilities of LLMs to speed up qualitative coding processes, utilizing OpenAI's language models GPT-3 and ChatGPT 3.5 in combination with a tailor-made codebook [16][29][31]. Tai et al. concluded that LLMs can be a useful aide as a validation tool to check the coding outputs of human coders. Moreover, deployment of a language model helped the researchers uncovering implicit assumptions in the codebook. In their paper, the authors highlighted the necessity of completing several runs per query – at least 10 and possibly more – to obtain stable results, as well as the effort that is needed to design context-relevant prompts [29]. These studies were conducted with non-open-source language models, which are trained on massive data corpora and therefore likely to yield better results, but are also somewhat problematic in an academic context, as we will discuss below. In addition to the deployment of LLMs at large in an academic context, researchers have started to explore the use of language models that have been trained on relatively small datasets – often abbreviated to SLMs or sLLMs – especially in fields with specific scientific or ethical needs, such as healthcare and law. Wang et al. published an extensive assessment of sLLMs that are used in these fields because of their relatively small computational needs, cost-effectiveness and adaptability to specific needs [30].

It has been pointed out by diverse researchers that there are several challenges with the application of machine-learning techniques to documents in a historical context. Nevertheless, large steps have already been taken in the field of Handwritten Text Recognition, which can historians save a lot of time in transforming hand-written records into manipulable datasets [6][20]. In general, however, historical text corpora are often too small to train a language model from scratch. Models

trained on modern data, however, tend to be biased because they make their interpretations from a modern, western perspective, since they are based on data originating from this time period and geographical area. Several researchers came up with the solution to use models that were trained on present-day data, but then fine-tuned with smaller historical data sets [18][23]. Still, this strategy requires the availability of thousands of data points for training purposes, implicating that to be worth the effort, one needs quite a large data corpus that the model can be applied to once it has been fine-tuned.

For this and other reasons – among them the tendency of historians to be not overly tech-oriented – there is also a group of researchers in the field of historical studies who do make use of language models that are trained on present-day data, being well aware of the limitations of these models when applied to historical data [17]. Chartier et al. conducted a comparative study to the capability of current LLMs to answer historical questions in the French context, without any fine-tuning or prompt-tweaking [7]. Their results varied considerably across language models, with one impressive forerunner (Gemini) and four to six promising next-bests. They found that the volume of the training set is not the main predictor of an LLM's performance in an historical context. Rather, the quality of the training data is decisive: in their case study, language models that were trained on academic sources, such as scientific articles and educational materials, performed better. They also suggested that the quality of the outputs would probably gain from designing good prompts and the modification of hypermetadata, such as temperature and sampling method. The approach we describe below is most similar to the yet unpublished work of Kaninik Baradi and Amineh Ghorbani at TU Delft [3]. With the help of language models, they assessed modern policy documents in the Indian city of Chennai to learn more about the equal distribution of climate adaptation measures. By making use of the Institutional Grammar framework [15], they tried to coax similar answers from the language model as we did: what type of rules do we encounter, to whom are these applicable, what are the actions that take place when a rule is triggered, etcetera. In a benchmark of several language models, they found an accuracy rate of around 66% on these tasks for the language model we used (see below). Comparison with other studies – taking into account that the methods and objectives are quite different – accuracy rates of over 70% (historic Chinese poetry classification with a fine-tuned model [18]; answering historical questions without any tuning [7]) should be within reach when using present-day LLMs to analyse historical data.

3. Methods

When selecting a language model for our experiment, two aspects were important: it had to be an open source model, and it had to be operable without the use of external computational power and with a minimum of technical expertise. The use of proprietary (s)LLMs as an aide in scientific research context is problematic, because it means that researchers lose control over their data and make concessions to the reproducibility of their work – which in our view runs counter to one of the basic principles of scientific research. Open source language models offer transparency regarding their architecture, algorithms, and underlying training data. If desired, the model can be adapted or fine-tuned to fit ones specific needs [18]. In addition, most open source models benefit from crowdsourcing networks: active user communities that inspect codes, make improvements, share knowledge and use cases. To prevent the unsolicited exposition of research data as a resource for

the training of future AI releases – which is especially important when working with sensitive data – the selected model should be able to operate in a sandbox environment, on either a stand-alone computer or a virtual machine.

The requirement to use a low-threshold language model – both in terms of computational power and background knowledge – has to do with the typical aptitude among historians. Although Digital Humanities is a growing field, the average historian is a tech-user rather than a tech-contributor. As a result, the budget reservations for IT resources are usually low – although that may change in the future – which points in the direction of an sLLM. Also, the number of colleagues who are used to adapt existing IT resources to fit their specific research needs is as yet limited. Therefore, humanities researchers need a safe and user-friendly coding aide to their disposition to convince them of the value of language model for their work. In a broader sense, we believe it is worthwhile to take stock of the current status of the application of machine learning techniques in social sciences and humanities, low-threshold sLLMs among them. Taking these considerations into account, we settled for the open source Microsoft Phi-3-mini 128k language model. Trained on 3.8 billion parameters – derived from publicly available educational materials, high-quality chat data, and synthetic textbook-style data [1] – it falls in the category of small Large Language Models. Although it cannot compete with LLMs that have been trained on massive data sets, such as ChatGPT and DeepSeek, its performance is nearing that of larger models such as Mistral, Gemma, and Llama [1]. Baradi and Ghorbani concluded in their benchmark that Phi-3 was performing somewhat worse than Mistral, Mixtral, Llama, and particularly Gemma, but still on a useful level [3]. On the flip side, the model can be run easily from an ordinary laptop computer or, according to Microsoft, even from a mobile phone. We operated Phi-3-mini within the Hugging Face ecosystem [19], which has a lively community sharing knowledge, expertise, as well as Python scripts that helped us on our way. To prevent data leakage, we made use of a virtual machine in a Google Colab environment.

For the sake of exploring the possibilities of the sLLM for qualitative data coding, we opted not to reiterate the entire coding process that was done during the Common Rules project. Rather, we simulated a basic version of the coding process to discover the model's potential, its strengths, weaknesses, and eventual pitfalls, based on a few specific features of rules which have also been identified in the Institutional Grammar (IG) literature [Crawford & Ostrom 1995]. Our approach thus tests the applicability of LLMs for both the Common Rules approach and the Institutional Grammar approach. This basic version can later be refined to meet the specific needs of a project. The coding process with which we tried the LLM contains four elements:

- Rule form (IG: Deontic): does the rule entail an obligation, prohibition, or permission? This element forms the backbone of the rule, defining the other elements of the rule's syntax.
- Subject (IG: Attribute): to whom does the rule apply?
- Action (IG: aIm): which action is expected when the rule is triggered?
- Object (IG: oBject): to what or whom is the action to be applied?

In the future, a larger set of elements will be taken into account, with additions such as the condition that triggers the rule and sanctions that are applied upon non-compliance. To clarify the use of the

aforementioned elements, we will take a look at how a rule from the Common Rules Database would have been coded manually for each of these elements (see also [21] and the codebook of [11]). In 1618, the clerk of the common Exel in the Netherlands recorded [11], rule 2483: "Furthermore, the commoners have approved the rule that anyone who is entitled to a share, based on the purchase of any land, will have to donate a barrel of beer." This rule applies to "anyone who is entitled to a share" (subject), is activated by "the purchase of any land" (condition), while the term "will have to" implies an obligation (rule form). The expected action is to "donate" (action) and the anticipated donation is "a barrel of beer" (object). In this rule, no sanction is imposed. In the original Common Rules project, several of these elements would have been condensed, while others were expanded further. For example, the codebook restricted the coding options for the subject of a rule to "everyone", "non-members", "members in general", "specific groups of members", "specific members", "officials in general", or "specific officials". The object "a barrel" would have been split in a measurement unit (barrel) and a quantity (1). Although it would – with quite some effort – probably be possible to request such a level of reduction from the LLM, this is beyond the scope of this paper.

To investigate the abilities of the LLM, we prepared a .csv-file containing the subset of rules from the Common Rules database that was based on Dutch cases. All records had been translated into modern English. Each record contained a unique rule identifier, the text of a single rule statement, its rule form, and the original codings regarding subject, condition, action, and object. Those records that took another rule form than either obligation, prohibition, or permission – such as the general statement "The aforementioned chairman of the common's assembly and the commoners stand firmly by the resolutions, previously taken" [11], rule 2873 – were removed from the data set. It also proved necessary to split several rules, because they appeared to contain more than one rule per statement after all. This was for instance the case in a rule issued in 1662, which is in fact the prohibition to export manure combined with an obligation of the officials to monitor this measure: "In case anyone is found to transport any manure from this common to other commons, he will have to pay a fine of 2 old ecus for each offence. The cattle pounders will have to see to it that this sanction will be executed" [11], rule 2603. The resulting data set contained 558 records. From the remaining data set, 72 records were selected at random, using a Python script. Out of these, 15 records were randomly selected to act as examples for the sLLM: 5 records that had originally been coded as obligations, 5 prohibitions, and 5 permissions. Upon their selection, we checked whether these examples were clear beyond doubt. Because of the rather neutral stance in the declaration "And furthermore it has been stated that the sheep will be counted after Saint-Lambert's Day" [11], rule 2690 – which can be read as either an obligation or a permission – this record was swapped for a clearer one. We used the remaining data set of 57 records as a working set to build a workable qualitative coding procedure that would perform as good as possible when categorizing the 4 elements mentioned above.

To test the general applicability of the qualitative coding procedure that we designed based on the working set, we also prepared 6 test sets. To be able to determine the impact of the heterogeneity of the data on the performance of the sLLM, we selected available datasets containing data from (a) different types of institutions for collective action (ICA), (b) different geographical regions in western Europe, and (c) different time periods. As a side-effect, the datasets differed highly in size and were not structured in the same systematic manner as the Common Rules database. The

following datasets were selected.

- The Dutch records of the Common Rules database. Type of ICA: mixed-resource common. Geographical area: central and eastern Netherlands. Time period: 12th-18th century. Number of records: 558. Note: this set includes the data used in the working set [11].
- Cofradías. Type of ICA: fishery cooperative. Geographical area: northern Spain. Time period: 16th-18th century. Number of records: 458. Reference: [26][27], coded by Grant T. Halliday BA, research intern at the SEICA research group, not yet published.
- Speulderbos common. Type of ICA: woodland common. Geographical area: central Netherlands. Time period: 17th-19th century. Number of records: 103. Reference: [22], collected by Ir. F. Kwadijk as part of ongoing PhD research, not yet published.
- Brewers' guild of Haarlem. Type of ICA: craft guild. Geographical area: western Netherlands. Time period: 17th century. Number of records: 43. Reference: [24], coded by dr. M. Groep-Foncke as part of ongoing research, not yet published.
- Coöperatie Eigen Hulp. Type of ICA: consumer cooperative. Geographical area: eastern Netherlands. Time period: 19th-20th century. Number of records: 48. Reference: [12], collected by drs. P. Steenbergen as part of his research to consumer cooperatives, see [28].
- SMART. Type of ICA: platform cooperative. Geographical area: Belgium. Time period: 21st century. Number of records: 236. Reference: [5], coded by dr. D. Bunders as part of his PhD research, see [4].

4. Results and discussion

The first task we undertook, was to find the parameters with which the LLM came up with the best results within the computational limits we had set. These parameters entailed the temperature, the number of beams, and the maximum of newly generated tokens, which we will now discuss. The "temperature" is a control mechanism that impacts the probabilities in a pre-trained language model: the higher the temperature, the more randomness in the answers the model generates. At a temperature of 0, the model will consider the token with the highest computated probability as the only possible outcome. By default, the temperature is set at 1, allowing the model some randomness in its answers. Setting a temperature lower than 1 is typically done in cases where it is important to stick close to the input text, either in content or in tone of voice. Temperatures of over 1 are often used when the aim is to generate creative outputs. As expected, the LLM went in overdrive when the temperature was set higher than 1, especially when requested to determine the rule form: nearly none of the rules were classified as either an obligation, permission, or prohibition. This resulted in a high number of the outcome 'nan' – in fact a way to communicate that no answer could be given. After 10 runs with each promising candidate of temperatures between 0 and 1, we settled for a rather conservative temperature of 0.3, which gave slightly better results than the standard temperature of 1.

The number of beams determines the focal length of the LLM. When predicting the next token, the model will by default look one step ahead at the probability of each possible next token to occur.

With each added beam, the model will look one step further, in the sense that it takes into account the probability of a sequence of tokens, which has the length of the number of beams. For instance, when predicting the next token after 'one is obliged to have', a one-beam search may yield 'a' as most probable next token. In a two-beam computation 'been' might be a more probable outcome, because it opens up more possibilities than 'a': 'been working', 'been present, 'been a', etcetera. Unfortunately, making variations with three or more beams often resulted in run-time errors, while there was barely a difference between the deployment of one or two beams. Hence, we decided to keep the number of beams at the standard, which is also known as a "greedy search" [19].

Although a too large increase in the maximum number of generated tokens may also lead to run-time errors, we found this strategy more promising. The number of generated tokens defines the maximum length of the answer the LLM comes up with. This is not important when the solution is expected to be one single word, such as obligation, permission, or prohibition. However, we will see later that this added bit of information does make a difference when trying determine the object or a subject of a sentence. In our case study, a maximum of twelve generated tokens proved the ideal setting, giving the language model sufficient room to come up with meaningful answers, but evading the risk of too many run-time errors.

Throughout our exercise to build a qualitative coding procedure, we utilized the same temperature, number of beams, and maximum number of generated tokens in every step. In future, it may be worthwhile to employ a different setting of the parameters per step in the process, adapted to the specific needs of the respective steps. For reproducibility, we set a seed parameter. This way, the text generating mechanism of the language model is initialized each time in the same manner, ensuring that the start position of the model to which we feed the text for next token prediction is reproducible. Needless to say that consecutive runs were operated with different seeds, or their outcomes would have been the same.

There were two reasons to start the design of the qualitative coding process with the detection of the rule form. As mentioned before, the rule form serves as the backbone of each rule, on which all other information is built. Moreover, it has a limited number of possible values, which should make the classification by the LLM a relatively easy task. To strengthen the performance of the language model, we decided to prime it in two different ways. First, we took a few-shot approach. This means that we fed the model five good examples of each possible outcome (obligation, permission, or prohibition), so that it could check each record against these representative examples to come up with the most probable solution. In addition, we incorporated some clues in the prompt that urged the model to consider certain keywords that point to a certain rule form, such as 'allowed' in case of a permission. After trying out several varieties, we came up with the following prompt as the most promising one.

```
"""A rule can comprise either a permission (often characterised by
words like 'authorized', 'allowed' or 'may'), an obligation (often
characterised by words like 'should', 'shall', 'must', or 'have to
be'), or a prohibition (often characterised by formulations with
'none', 'no one', 'nobody', or 'forbidden'). Based on the following
texts and labels:\n\n""" + '\n'.join(('Text:' +
stratified_examples['text_translation_en'] + '\nLabel: ' +
```

```
stratified_examples['rule_form']) + '\n') + """ What kind of content
does the following rule comprise? When assessing the content of the
rule, ignore any sanction mentioned that will apply only when the
rule you're assessing is breached. You can only answer with
'permission', 'obligation', or 'prohibition'. When the main
statement in the rule seems to be neutral, choose 'obligation'."""
```

The addition to ignore any sanctions mentioned in the text, proved necessary because the model sometimes had the tendency to confuse the initial rule with the sanction – treating the latter as a rule in its own right. For clarification, we will look at the rule 'Regarding the collecting of peat, it has been decided that peasants who own a horse will be allowed to dig for half a day, and those not owning a horse will be allowed to dig for a day; in case anyone digs above this amount, he will be sanctioned as usual. In case anyone may already have dug any peat not in accordance with this regulation, the digging will have been done in vain and the peat will revert to the common' [11], rule 2617. Without the prompt to ignore any sanctions, the rule was consequently categorized as a prohibition. After adding the said message to the prompt, the categorization improved: in all runs but one, the rule was categorized as a permission. The addition to the prompt to categorize apparently neutral statements as obligations should only be made after due consideration. In the case of our working set, inclusion of the statement improved the categorization considerately. However, obligations make up the majority of the records in this particular data set. This means that if categorization as an obligation is implicitly preferred, it is likely to result in an improvement of the overall outcomes. In our case, correct identification of obligations rose from less than 50% to over 90%, while the identification of prohibitions sank from around 90% to around 80%. The biggest loser was the category "permission", which correct identification dropped from 85% to around 60%. However, only 3 out of 18 permissions (17%) were categorized incorrectly in all 5 runs. This suggests that the (automated) comparison of several runs will result in the placement of the necessary red flags.

When we broaden the scope from the working set to the test sets, we can conclude that obligations are the most-occurring rule form, the least so in the 21st-century SMART data set (111 obligations, 103 permissions, 22 prohibitions) [5]. This means that we may reasonably run the risk of over-fitting by favoring obligations over other rule forms for time-saving reasons, provided that we are aware of the limitations. One possible strategy would be to do two or more series of runs with diversified prompts – with and without favoring obligations – and have the outcomes compared. Applied to our working set, this led to less than 4% incorrectly categorized records. However, it also meant that 44% of the outcomes had to be checked by humans. A time-saving strategy could be to check those outcomes that were internally inconsistent after multiple runs with the same prompt, including the urge to consider apparently neutral statements as an obligation. In our case, after 5 runs, we would have been forced to check 7% of the outcomes. However, the accuracy rate dropped somewhat from around 96% to 91%. The next steps was to determine the subject, action, and object of each record: to whom does the rule apply, which action is expected, and to what or whom is this action applied? Because the detection of the subject proved the hardest task to the LLM, we decided to gather the action and the object first, and then use the aggregated information to determine the subject. For all of these three steps, we tried two different approaches.

The first alley we tried was grammar parsing, telling the language model that it was acting as a linguist:

```
"""You are a linguist doing syntax analysis on complex sentences.
\n\nGiven the verbal statement """ + categorize_ICAs_rules['rule_verb']
+ " to " + categorize_ICAs_rules['action_verb'] + """, what is the
object of the following statement?"""
```

Even after several instances of prompt tweaking, the grammar-based approach obtained a meager accuracy rate of between 36-54%. The solutions the language model came up with seem to have been chosen at random, as can be shown by the deconstruction of the following example. 'Since there have been complaints that the common land of this village is being diminished by the driving of sheep, belonging to servants, onto the common, it has been devised that no servant will be allowed to keep any sheep anymore, each offense to be fined at a daalder' [11], rule 2534. The LLM put the rule form correctly down as a prohibition. Subsequently, it chose 'drive' as the action prohibited – which is near the mark, although 'keep' would have been better. Finally, it registered 'common land' as the main object and 'servants keeping sheep' as the subject, which would result in the simplified rule 'servants keeping sheep are prohibited to drive common land.' In a sense, it is odd that the LLM encounters so many difficulties with a grammar-based approach, because other results – when the model was not asked to take the syntax into consideration – show that the model does look at the way that sentences are usually formed. The rule 'prohibition to any tenant farmer (. . .) to mow the grass (. . .)' [11], rule 2935, was simplified by the model to 'people are prohibited to any tenant farmer.' Similarly, 'obligation to all, peasants as well as farmers, to perform the maintenance of the small leat (. . .)' [11], rule 3012, was deconstructed to 'people are obligated to all, peasants as well.' In both these cases, the word 'to' is apparently taken as a clue that the next token will be a verb, so that 'any' and 'all' respectively are noted down as the expected action. The second approach we tried was to have the LLM complete a sentence, building on the information gathered before. Thus, the rule form was used to determine the expected action. The combination of the rule form and the action led the LLM to the object of this action.

```
"""Based on the following texts, complete the following sentence:
\n\n People are """ + categorize_ICAs['rule_verb'] + " to " +
categorize_ICAs['action_verb'] + """... Do only mention the object
that replaces the '...', without repeating other parts of the
sentence."""
```

When the aforementioned steps were run consecutively without intermediate checks, the accuracy rate of the three steps combined lay between 70-80% throughout the working set and test sets. In any case, it proved important to give the language model sufficient instructions to do the task. The instruction not to repeat parts of the sentence was necessary to refrain the model from giving too verbose responses, thereby not coming to the core of the message before the maximum number of generated tokens was reached: 'sheep keepers are prohibited to keep...those found.' In the step where we required the LLM to detect the action, we added the instruction to start the answer with a verb. Thus, we could easily isolate the verb with a Python script to utilize that for the

detection of the object in the next step. Sometimes, the best possible prompt was found in devious ways. As mentioned before, the most difficult task for the LLM to perform was to determine the subject: to whom does the rule in question apply? Therefore, we abandoned our previous attempts to detect the subject immediately after the rule form, before looking at the action and the object. Hence, we could no longer use the format '[subject] is or are [obligated/permitted/prohibited] to ...' to find the action or object. As a work-around, we came up with the format 'people are [obligated/permitted/prohibited] to ...' Possibly because of the explicit reference to people, the language model performed better, even compared to the step in which we used a corrected list of subjects as a starting point for action detection. After the rule form, action, and object had been established this way, the determining of the subject also went better than in any previous attempt.

When searching for the subject, we explicitly gave the instruction to put down rules that had apparently no subject as being applicable to everybody. Although this measure seems comparable to choosing 'obligation' as the default rule form – as discussed above – we believe this is something different. Most human coders would understand without being told that legislation is generally applicable, unless otherwise stated in the rule text. Conversely, a language model has to be provided with this background information, lest it starts looking in the text for a possible candidate – trying its best to come up with an answer. Generally applicable prohibitions come with another problem. They are often phrased as a negative sentence, like this one from the cofradías-data set: 'and furthermore, we order and mandate that no commoner can ask for money from the captains after the season in which they have earned them has passed' [26], rule 52. The sLLM correctly categorized this rule as a prohibition, but came to the conclusion that 'nobody is prohibited to request money from the captains.' Giving the model the instruction to consider phrases with 'nobody' and its synonyms as applicable to everybody yielded meager results. Fortunately, this error could be easily dealt with by changing all instances of 'nobody' into 'everybody' with a simple script.

A more profound problem is the fact that the word 'everybody' is sometimes literally mentioned in the text, although it is bound to refer to the entire community of commoners rather than to everybody in a literal sense. One could argue, for instance, that the statement that 'everybody should have their peat transported out of the moorlands before Saint-Michael's Day' [11], rule 2497, is only applicable to those who have obtained the right to dig peat in the first place – and thus to all members of the common. This kind of issues makes us aware of the interpretative choices we make. They should be addressed in the codebook, regardless of working with human or machine-aided coders. Moreover, one should consider the option of letting both codings – a grammar-based and an interpretative one – coexist with proper annotation.

When applying the qualitative coding process – designed for the early modern, Dutch working set of commons' rules – to the six test sets from different sectors, time periods, and regions, we found similar accuracy rates in four out of six test sets: 70-80% for the combined detection of rule form, action, and object without intermediate checks; and 70-75% for the sole detection of the subject. When compared to the findings of Baradi and Ghorbani, this performance is higher than they found for Phi-3, comparable to their results with Mistral/Mixtral, and lower than those of Llama and Gemma [3]. Manjavacas and Fonteyn observed that language models trained or fine-tuned on historical data performed less well on relatively young (post-1800) materials [23], our prompt-based approach seems to work as well for medieval records as for an early 20th-century consumer cooperative and a modern platform cooperative. This does make sense: considering the

biased performance towards historical data of models trained on modern data, a model trained or fine-tuned on historical data is bound to perform imperfectly when applied to more recent data. We used a sLLM trained on present-day data, opting for tweaking the prompt to fit our specific needs without altering the training data. The least-performing among our test sets were those of the cofradías and of the brewers' guild. On average, the cofradías' data set yielded 69% accuracy on the rule form-action-object combination (and a surprising 84% correctly identified subjects), while the guild's results did not exceed 62%. On closer inspection, a series of peculiar reversals stood out: 'people are obligated to refrain from walking off their services' [24], rule 8, where a prohibition leave one's employer after a certain date was meant; 'people are prohibited to ignore the order that ships who return' [26], rule 10, instead of the obligation to await one's turn. These errors are probably due to the fact that we did not check the outcomes in between steps, but decided to determine the rule form, action, and object in one go. Once the language model had made a mistake in assigning the rule form, it chose a verb that reversed the erroneous rule form, in order to get the action right.

Some positive surprises were encountered in the outputs of the larger Common Rules and the Speulderbos data sets[11][22]. It turned out that one of the entry records of the Common Rules had not been translated into English after all, but was still in its original early modern Dutch language, which differs considerately from modern Dutch. Nevertheless, the LLM succeeded in identifying the rule form, action, and object correctly. Although Phi-3 was trained predominantly on American-English texts, supplemented with text corpora as wide apart as Russian, Arabic, and Vietnamese [1], its ability of detecting the structure and contexts of written texts is admirable. The good results of the Speulderbos coding process (>73% accuracy) were remarkable because the input text is in no way structured as a rule set, but rather as a collection of oaths and verbal meeting reports. Apparently, the LLM was able to read between the lines and deduct rules. Still, any researcher working with language models – and historians in particular – has to be well aware of the model's limitations. The historical context complicates the qualitative coding process with ready-made (s)LLMs. In our case, four rules of the working set entailed an obligation to ring the pigs before they were released on the common. Nowadays, it is no longer practice to provide pigs with a nose ring to prevent them from spoiling the soil by their rooting behavior – and therefore it was probably not much debated in the training data used for Phi-3. Hence, the model had considerate difficulties with these records, unless the term 'nose ring' was explicitly mentioned. The fact that the language model functions as a next token predictor shone through in the few instances where it came up with the solution that people were obliged to ring the bells – an understandable mistake when taking into account that the model was trained on modern data.

5. Conclusion and future work

The deployment of the current and imminent generation of LLMs as an aide to onerous tasks such as qualitative coding is a promising alley to explore further for researchers in the fields of Social Sciences and Humanities. These aides are not ready for use of the shelve, in the sense that it takes some effort to find the right prompts and hypermetadata that are fitting to the job at hand. However, when the availability of training data is limited – which is often the case in historical research – there is no pressing need to train language models from scratch or go through

the process of fine-tuning an existing model. Even with a sLLM such as Phi-3, it is already possible to obtain reasonable results once the process is set up well. To use (s)LLMs in historical and other Social Sciences and Humanities studies successfully, one needs to invest time in finding the most effective settings, including hypermetadata, the sequence of tasks assigned to the language model, and the right prompt. When doing so, it is worthwile to try out and combine several strategies, comparing one-shot and few-shot approaches, including sufficient background information or clues in the prompt, and requesting the model to disregard certain information. Setting up the process well takes time, but in the end it pays off. That said, researchers should continuously be aware of the limitations of the tools they use. In the case of (s)LLMs, this means having a good understanding of the type of data that the model was trained on and its strengths and weaknesses on specific tasks, including a benchmark towards other models. This is especially important when applying a ready-made language model – which was probably trained on modern data – in a historical context, but may also be an issue when the context changes in another way, for instance to an entirely different geographical region or cultural context. Information on training data and benchmarks can be found in the language model's release documentation, as well as among user communities such as the Hugging Face ecosystem.

For future research, we want to propose several directions. First, we suggest to replicate the research done so far – and the extensions described below – using other open source LLMs, such as Llama, Mistral/Mixtral, Gemini/Gemma, but also the newly released Phi-4-mini model, even though some of these may require a larger amount of computational resources. We assume that with the utilization of Phi-3 only the baseline of what is possible has been found. Building further on the expertise gained in this process, we would like to turn to the detection of somewhat more complicated elements, such as the condition that triggers a rule or the sanction imposed upon non-compliance. The first attempts made in this direction by Baradi and Ghorbani seem hopeful [3]. Going even one step further, we want to explore to what extend LLMs are capable of detecting more abstract concepts, such as the presence of certain knowledge that forms the starting point of rule-making processes.

Acknowledgments

This publication is part of the project Building a UNified theory for the development and resilience of Institutions for Collective Action for Europe in the past millennium (UNICA) with project number VI.C.191.052 which is financed by the Dutch Research Council (NWO).

References

[1] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A.A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, A. Benhaim, M. Bilenko, J. Bjorck, S. Bubeck, M. Cai, Q. Cai, V. Chaudhary, D. Chen, D. Chen, W. Chen, Y.-C. Chen, Y.-L. Chen, H. Cheng, P. Chopra, X. Dai, M. Dixon, R. Eldan, V. Fragoso, J. Gao, M. Gao, M. Gao, A. Garg, A. Del Giorno, A. Goswami, S. Gunasekar, E. Haider, J. Hao, R.J. Hewett, W. Hu, J. Huynh, D. Iter, S.A. Jacobs, M. Javaheripi, X. Jin, N. Karampatziakis, P. Kauffmann, M. Khademi, D. Kim, Y.J. Kim, L. Kurilenko, J.R. Lee, Y.T. Lee, Y. Li, Y. Li, C. Liang, L. Liden, X. Lin, Z. Lin, C. Liu, L. Liu,

M. Liu, W. Liu, X. Liu, C. Luo, P. Madan, A. Mahmoudzadeh, D. Majercak, M. Mazzola, C. César T. Mendes, A. Mitra, H. Modi, A. Nguyen, B. Norick, B. Patra, D. Perez-Becker, T. Portet, R. Pryzant, H. Qin, M. Radmilac, L. Ren, G. de Rosa, C. Rosset, S. Roy, O. Ruwase, O. Saarikivi, A. Saied, A. Salim, M. Santacroce, S. Shah, N. Shang, H. Sharma, Y. Shen, S. Shukla, X. Song, M. Tanaka, A. Tupini, P. Vaddamanu, C. Wang, G. Wang, L. Wang, S. Wang, X. Wang, Y. Wang, R. Ward, W. Wen, P. Witte, H. Wu, X. Wu, M. Wyatt, B. Xiao, C. Xu, J. Xu, W. Xu, J. Xue, S. Yadav, F. Yang, J. Yang, Y. Yang, Z. Yang, D. Yu, L. Yuan, C. Zhang, C. Zhang, J. Zhang, L.L. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, X. Zhou, *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*, retrieved from ArXiv (2024) arxiv.org/pdf/2404.14219

[2] M. Ale Ebrahim Dehkordi, A. Ghorbani, G. Bravo, M. Farjam, R. van Weeren, A. Forsman, T. De Moor, *Long-Term Dynamics of Institutions: Using ABM as a Complementary Tool to Support Theory Development in Historical Studies*, *Journal of Artificial Societies and Social Simulation* **24:4** (2021) 7 doi 10.18564/jasss.4706

[3] K. Baradi, A. Ghorbani, *Ex-ante measurement of procedural justice in climate change adaptation. Application of the Institutional analysis approach to assessing alignment with climate justice*, presentation for Institutional Grammar Research Initiative 5 November 2024 youtube.com/watch?v=wICdIJxMmBo

[4] D.J. Bunders, *Gigs of their own: can platform cooperatives become resilient?* Erasmus University Rotterdam (2024) pure.eur.nl/files/136225935/

[5] D. Bunders, T. De Moor, *Institutional grammar coded data of rules in Smart, 2017-2022*, Dataset, EUR Repository no. 22717171

[6] C. Capurro, V. Provatorova, E. Kanoulas, *Experimenting with Training a Neural Network in Transkribus to Recognise Text in a Multilingual and Multi-Authored Manuscript Collection*, *Heritage* **6:12** (2023) 7482-7494 doi 10.3390/heritage6120392

[7] M. Chartier, N. Dakkoune, G. Bourgeois, S. Jean, *HiBenchLLM: Historical Inquiry Benchmarking for Large Language Models*, *Data & Knowledge Engineering* **156** (2025) 102383 doi 10.1016/j.datak.2024.102383

[8] S.E. Crawford, E. Ostrom, *A grammar of institutions*, *American political science review* **89:3** (1995) 582–600 doi 10.2307/2082975

[9] T. De Moor, *Revealing historical resilience*, *Science* **362/6420** (2018) 1238 sciencemag.org/content/362/6420/1236

[10] T. De Moor, M. Farjam, R. Van Weeren, G. Bravo, A. Forsman, A. Ghorbani, M.A. Ebrahim Dehkordi, *Taking sanctioning seriously: The impact of sanctions on the resilience of historical commons in Europe*, *Journal of Rural Studies* **87** (2021) 181-188 doi 10.1016/j.jrurstud.2021.08.009

[11] T. De Moor, R. van Weeren, J. Zuijderduijn, J.M. Lana-Berasaín, G. Sanz Lafuente, M. Bogino Larrambebere, A. Winchester, J. Bowen, *Common Rules. The regulation of institutions for managing commons in Europe, 1100 - 1800*, Dataset and codebook (2021) EUR repository no. 164585401

[12] *Eigen Hulp U.A. Coöp winkelvereeniging van Rooms Katholieken*, Gelders Archief, 1176 Handelsregister Arnhem 1921-1969, inventory number 4044

[13] M. Farjam, T. De Moor T., R. van Weeren, G. Bravo, A. Forsman, A. Ghorbani, M. Ale Ebrahim Dehkordi, *Shared patterns in long-term dynamics of commons as institutions for collective action*, *International Journal of the Commons* **14:1** (2020) 78-90 doi 10.5334/ijc.959

[14] A. Forsman, T. De Moor, R. van Weeren, G. Bravo A. Ghorbani, M. Ale Ebrahim Dehkordi, M. Farjam, *Eco-evolutionary perspectives on emergence, dispersion and dissolution of historical Dutch commons*, *PLoSONE* **15:7** (2020) doi 10.1371/journal.pone.0236471

[15] C.K. Frantz, S. Siddiki, *Institutional Grammar. Foundations and Applications for Institutional Analysis*, Palgrave Macmillan (2022)

[16] F. Gilardi, M. Alizadeh, M. Kubli, *ChatGPT outperforms crowd workers for text-annotation tasks*, *Proceedings of the National Academy of Sciences of the United States of America* **120:30** (2023) e2305016120 doi 10.1073/pnas.2305016120

[17] G. Gonzalez Garcia, C. Weilbach, *If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History*, in proceedings of *Computational Humanities Research Conference 2023* (2023) arxiv.org/pdf/2310.10808

[18] J. Hou, S. Zhang, *Exploring Thematic Diversity in Classical Chinese Poetry: A Novel Dataset and a BERT-enhanced Ensemble Learning Approach*, *Journal on Computing and Cultural Heritage* **17:4** (2024) 60 doi 10.1145/3685679

[19] Z. Hussain, M. Binz, R. Mata, D.U. Wulff, *A tutorial on open-source large language models for behavioral science*, *Behavior Methods Research* **56**(2024) 8214–8237 doi 10.3758/s13428-024-02455-8

[20] M. Koolen, R. Hoekstra, J. Oddens, R. Sluijter, R. Van Koert, G. Brouwer, H. Brugman, *The Value of Preexisting Structures for Digital Access: Modelling the Resolutions of the Dutch States General*, *Journal of Computing and Cultural Heritage* **16:1** (2023) 1-24 doi 10.1145/3575864

[21] M. Laborda Pemán, T. De Moor, *History and the commons: A necessary conversation*, *International Journal of the Commons* **10:2** (2016) 516-528 doi 10.18352/ijc.769

[22] *Malen wetten, rechten, gebruicken overgiften van Speulderbos, in dit nieuwe boeck overgeschreven in Julio des jaers Duysent ses hondert seventigh*, Gelders Archief, 366 Marken en Maalschappen, inventory number 726

[23] E. Manjavacas, L. Fonteyn, *Adapting vs. Pre-Training Language Models for Historical Languages*, *Journal of Data Mining and Digital Humanities* (2022) 9152 doi 10.46298/jdmdh.9152

[24] *Ordonnantie op de brouwerij, 1643*, Noord-Hollands Archief, 3993 Stadsbestuur van Haarlem, inventory number 3997

[25] E. Ostrom, *Governing the Commons. The evolution of Institutions for Collective Action*, Cambridge University Press (1990) doi 10.1017/CBO9780511807763

[26] P.A. Porras Arboledas, *Ordenanzas de la cofradía de mareantes de San Pedro y San Andrés de Comillas (1522-1662)*, *e-Legal History Review* **27:Extra** (2018)

[27] E. Prat, E. Serna, P. Vila, F. Rahola, *Les ordinacions de la pesquera de Cadaqués (s. XVI-XVIII)*, Palmós (2006)

[28] P. Steenbergen, *Van ons, voor ons. 150 jaar veerkracht van consumentencoöperatie Coop*, Fagus (2024) link

[29] R.H. Tai, L.R. Bentley, X. Xia, J.M. Sitt, S.C. Fankhauser, A.M. Chicas-Mosier, B.G. Monteith, *An Examination of the Use of Large Language Models to Aid Analysis of Textual Data*, *International Journal of Qualitative Methods* **23** (2024) doi 10.1177/16094069241231168

[30] F. Wang, Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, W. Wang, R. Li, J. Xu, X. Tang, Q. He, Y. Ma, M. Huang, S. Wang, *A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness*, retrieved from ArXiv (2024) doi 10.48550/arXiv.2411.03350

[31] Z. Xiao, X. Yuan, Q.V. Liao, R. Abdelghani, P.-Y. Oudeyer, *Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding*, in proceedings of the *28th International Conference on Intelligent User Interfaces* (2023) doi 10.1145/3581754.3584136