Uncovering Work from Words: LLM-Based Information Extraction from Historical Petitions

Ellinor Lindqvist, Eva Pettersson, Joakim Nivre
Uppsala University
Dept. of Linguistics and Philology
firstname.lastname@lingfil.uu.se

Abstract

We investigate the extraction and normalisation of phrases describing work from 18th-century Swedish petitions using four LLMs: GPT-4o, Llama-3 70B/8B, and Mixtral-8x7B. Performance is evaluated across four configurations: isolated extraction, isolated normalisation, a staged pipeline, and a combined multitasking setup, using both full and filtered texts (with formal greetings and closing sections removed). While exact phrase matching remains low ($F1 < .10$), token-level and semantic similarity scores suggest that models consistently locate relevant topical regions. Semantic similarity scores must however be interpreted with caution, since they are often only marginally higher than an average baseline. Results reveal a "multitasking paradox": combined extraction and normalisation improves phrase location for high-parameter models but degrades normalisation precision. Furthermore, normalisation benefits from the context of a staged pipeline compared to isolated tasks, while text filtering has only marginal effects. Despite a tendency towards over-prediction, qualitative analysis suggests that models can detect plausible work-related expressions missed by human annotators. These findings illustrate the challenges of historical extraction and suggest that hybrid human–machine workflows are a promising approach for enhancing coverage and interpretability in cultural heritage research.

Keywords: Information Extraction, Petitions, Large Language Models, Historical Swedish, Historical NLP, Digital Humanities, Text Normalisation

1. Introduction

The expanding field of digital humanities increasingly relies on advanced computational methods to analyse digitised historical sources. Although such sources are becoming more accessible, they remain challenging for automated processing due to linguistic variation, inconsistent orthography, and specialised historical contexts. Large language models (LLMs) show strong capabilities for complex language understanding, yet their suitability for domain-specific information extraction in structured historical texts, such as petitions, is still underexplored. In contrast to traditional encoder-based transformer models (e.g., BERT), which typically require extensive task-specific fine-tuning on annotated data, LLMs provide a more flexible few-shot alternative. This is especially beneficial in historical research, where large-scale labeled datasets are often scarce, enabling complex information extraction without the need for substantial model training.

This paper evaluates the ability of LLMs to extract and normalise phrases describing working activities in historical Swedish petitions. Extraction identifies where work is mentioned, while normalisation produces modernised and standardised forms of historically variable expressions. These normalised forms can support historians and digital humanists to analyse work and occupations across many documents and time periods. Automatically identifying information about how individuals supported themselves offers valuable insights into economic life and social relations in the past. We draw on annotated petitions from the Gender and Work (GaW) project at Uppsala University,[1] using a subset of this resource to test extraction and normalisation methods and to propose robust evaluation approaches for natural language processing (NLP) applied to historical texts.

Furthermore, early modern European petitions followed a classical rhetorical structure, offering cues that may assist automated analysis. As a secondary objective, we therefore examine whether rhetorical segmentation, for example removing the opening section (*Salutatio*) and the closing section (*Conclusio*), can enhance extraction performance.

Hence, our research objectives are:

1. Evaluate LLM performance in extracting phrases that describe work from Swedish historical petitions under different task settings.
2. Assess the impact of rhetorical segmentation by comparing full-text input with filtered versions.

This study contributes to our understanding of how LLMs can support historical research, demonstrating their potential and limitations for fine-grained information extraction in low-resource, structurally diverse corpora. We provide insights for digital humanities and NLP research on applying LLMs to complex historical documents and enabling more scalable analysis of historical texts.

[1] https://gaw.hist.uu.se/petitions/

Proceedings of LT4HALA 2026 @ LREC 2026, pages 19–37.
11 May 2026.

2. Background

Digital methods have become increasingly central to the study of historical sources, enabling large-scale discovery and analysis of cultural heritage materials that would be difficult to achieve manually (Wilson, 2022; Burrows, 2023; Wijfjes, 2017). Yet historical texts pose specific challenges for NLP, including substantial orthographic variation, diachronic linguistic change, and limited standardised resources (Piotrowski, 2012; Bollmann, 2019). Annotated training data for domain-specific tasks are also scarce and expensive to produce, limiting the feasibility of supervised adaptation or fine-tuning in many archival contexts (Krusic, 2024; Hiltmann et al., 2025).

The emergence of LLMs has opened new possibilities in digital humanities. For example, LLMs have been applied to detect irony in 19th-century Spanish newspapers (Cohen et al., 2025) or to transcribe handwritten archival documents more quickly and accurately than traditional HTR systems (Humphries et al., 2025). While these models offer strong zero- and few-shot capabilities, especially valuable in low-resource historical settings, they also introduce the risk of hallucination and domain-inaccurate outputs. For instance, Tudor et al. (2025) show that LLMs may misclassify or invent entities in zero-shot named entity recognition on historical text, whereas Hiltmann et al. (2025) demonstrate that domain-aware prompting can reduce such errors and better align model predictions with historical conventions.

Some historical documents, such as petitions, also exhibit structured rhetorical organisation that can support automated analysis. In much of premodern Europe, petitions followed a classical rhetorical structure, typically comprising five or six sections (Hansson, 1988; Sokoll, 2006; Israelsson, 2016):

1. *Salutatio*: Formal salutation to the addressee.
2. *Exordium*: Brief appeal to authority, goodwill, or capacity to help.
3. *Narratio*: Background and supporting reasoning, often including (4. *Argumentatio*).
5. *Petitio*: Explicit request or plea being made.
6. *Conclusio*: Closing courtesies, often including a signature.

Recent work on rhetorical annotation of 18th-century Swedish petitions (Lindqvist et al., 2025) shows that LLMs can reliably identify the *Salutatio*, *Petitio*, and *Conclusio*, with F-scores often in the high 90s. Such automated segmentation supports focused analysis of specific parts of historical documents, which could potentially make information extraction tasks more precise.

However, it remains uncertain whether rhetorical segmentation provides measurable benefits for downstream extraction tasks. In this work, our primary focus is on assessing how effectively LLMs can identify and normalise phrases describing work in historical petitions, while also exploring segmentation as a secondary factor that may influence performance.

3. Data

3.1. The Gender and Work Project

This study uses material from the Gender and Work (GaW) project at Uppsala University,[2] which investigates how people supported themselves in Early Modern Sweden (c. 1550–1800). The project gathers evidence of work-related activities from diverse archival sources, such as court records, account books, and personal letters, and stores the information in a research database. Each entry consists of a transcribed source document and metadata describing who performed an activity, what was done, when and where it occurred, and in what context, along with archival source information and, in many cases, images of the original documents to support historical transparency and traceability.

A central part of the project's method is the verb-oriented approach (Fiebranz et al., 2011), in which historians identify and extract expressions that describe work-related activities. These activities are often, but not exclusively, realised as verb phrases (e.g., "sell clothes", "fish herring").

Because work-related activities are context-dependent and not always expressed through standard verb-phrase structures (e.g., a verb and its object), the resulting annotations sometimes include fragmentary expressions, prepositional phrases, or nominal structures. This reflects the historical realities of expression rather than prescriptive grammar. To standardise interpretation for analysis, GaW annotators manually reformulate these expressions into modern Swedish infinitive forms (Ågren et al., 2023). Table 1 provides examples of this process, illustrating the shift from archaic, historical phrases to normalised, modern verb phrases, often using an infinitive verb followed by a direct object.

For computational processing, however, this creates a non-trivial modelling challenge. Work activities must be detected not only at the lexical level, but also semantically and contextually: for instance, the Swedish verb "köpa" (*to buy*) may indicate commercial trade in one context but personal consumption in another. Combined with substantial orthographic variation in Early Modern Swedish, this makes the GaW data both historically rich and

[2] https://www.uu.se/en/research/gender-and-work

Historical Phrase	Manual Normalisation
I lånebanquen à 4 1/2 proc:t *(In the Loan Bank at 4 1/2 percent)*	låna ut pengar till bank *(to lend money to a bank)*
Förteckning på mitt husfolck *(Register of my household members)*	lämna mantalsuppgift *(to submit a census return)*
blifwit föranlaten at söma åt andra *(been obliged to sew for others)*	sy kläder *(to sew clothes)*
Till bränuinstillwärkning åtager iag mig *(For liquor production I undertake)*	bränna brännvin *(to distill liquor)*
inbrotts stöld *(burglary theft)*	göra inbrott *(to commit burglary)*

Table 1: Examples of manual normalisation by GaW historians. English (approximate) translations are in italics.

Origin	Time Period	#Docs	#Phrases
Stockholm	1667–1734	57	340
Västmanland	1758	9	109
Örebro	1719–1800	44	84

Table 2: Test data statistics, including the number of documents and annotated phrases that describe work per collection.

valuable for research, while presenting complex challenges for automated analysis.

3.2. The Petition Dataset

This paper focuses on 18th-century Swedish petitions drawn from the GaW database (Section 3.1). The material comes from three collections: petitions to the County Administrative Boards of Västmanland and Örebro, and petitions to the Board of Trade in Stockholm. Two documents used during prompt development are excluded from evaluation. The final dataset comprises 110 petitions, with detailed statistics shown in Table 2.

4. Method

We evaluate LLM-based extraction of phrases describing work from historical petitions under two main conditions: using the full transcription and using a version where selected rhetorical sections are removed. Below, we describe the filtering procedure, extraction tasks, prompt design, and evaluation framework.

4.1. Rhetorical Filtering

To assess whether removing greetings and closing statements improves extraction performance, we create a filtered dataset where the *Salutatio* and *Conclusio* are removed. Since these sections predominantly contain formulaic courtesy expressions, we hypothesise that they rarely include work-related information.

Rhetorical segmentation is performed using Llama 3.3 70B (Touvron et al., 2023) following the methodology of Lindqvist et al. (2025). To ensure complete coverage, detected spans are projected back onto the original text using semi-soft alignment: for each annotated segment, we perform a case- and whitespace-insensitive search for the closest matching occurrence and extract the corresponding substring. This procedure successfully mapped the *Salutatio* in 96% and the *Conclusio* in 88% of the documents.

4.2. Models and Extraction Tasks

We evaluate four models selected to represent a diverse range of sizes, architectures, and licensing: two LLaMA open-weight transformers from Meta AI of different sizes, Llama 3.3 70B and Llama 3.1 8B (Touvron et al., 2023); the open-source model Mixtral 8x22B (Jiang et al., 2024) from Mistral AI; and GPT-4o from OpenAI (OpenAI, 2026). All models are accessed via APIs using identical prompts and inference settings across runs. All experiments use temperature 0 for deterministic and reproducible outputs. Four tasks are evaluated across both full and filtered datasets:

1. **Extraction:** Identify phrases describing work from the historical source text, preserving original wording.
2. **Isolated Normalisation:** Given gold phrases manually extracted by historians, normalise them to modern Swedish infinitive forms.
3. **Pipeline Normalisation:** Normalise phrases extracted by the model from the source text.
4. **Combined Extraction and Normalisation:** Simultaneously identify historical phrases and provide normalised equivalents.

You are an expert in linguistic analysis of Swedish historical texts. Your task is to analyse Swedish petitions from the 18th century and identify all phrases describing work, work performance, or work-related actions.

 # Main Instructions

1. **Read the text carefully.** The text may contain archaic spelling, free word order, and compound words.
2. **Extract only phrases describing work, work performance, or work-related actions.** This includes performing, contributing to, or participating in work, occupations, service, or duty for a livelihood, regardless of whether payment is given.
3. **Preserve the original text exactly.** Do NOT add, remove, or change words, spelling, or punctuation.
4. **Return a list of extracted phrases** in the order they appear in the text.

 # Output Format

Return only a JSON object with the key `"extracted_phrases"` and a list of phrases from the text. Return JSON exactly in this format, without extra text, comments, or code-block markers:

```
{ "extracted_phrases": ["grind some grain", "in service have been"] }
```

 # Examples

(Only short excerpts are shown in the examples. In the task, the entire text should be processed.)

Input 1: "/. . . / since at this time hops were so hard to be come by? Was answered she had bought hops from Tidö, whereto she had made a small addition of some Wormwood Sprigs, /. . . /"
Output 1: `{ "extracted_phrases": ["purchased hops"] }`

/.../

 # Identify and extract the phrases describing work in the following petition:

Figure 1: The prompt used for Task 1, translated to English, including role definition, extraction instructions, formatting constraints, and few-shot examples.

4.3. Prompting and Data Recovery

The pipeline uses few-shot prompting with twelve example snippets from the GaW corpus. All prompts follow a consistent structure: (1) role instruction positioning the model as an expert in linguistic analysis of Swedish historical texts, (2) explicit instructions emphasizing preservation of original features (archaic spelling, word order), (3) clear definition of phrases describing work as expressions describing performance of, contribution to, or participation in work for earning a living, and (4) step-by-step task instructions with examples. This shared structure is exemplified by the Task 1 prompt in Figure 1 (abbreviated to one of twelve few-shot examples), and all prompts and their English translations can be found in the Appendix. Output is requested in structured JSON.

To address API timeouts and parsing failures, we implemented a three-step recovery protocol. First, models were rerun on the specific input documents that previously yielded empty or failed outputs (up to two additional attempts). Second, for model outputs remaining unparseable, an *ad-hoc* regex-based script salvaged phrases from malformed or truncated JSON and pruned repetitive loops by collapsing phrase repetitions into single instances.

Model	#Docs	R1	+R	+Rep
GPT-4o	880	94.9%	98.7%	100%
Llama 70B	880	99.4%	99.5%	100%
Llama 8B	880	72.1%	73.1%	99.8%
Mixtral 8x22B	880	66.7%	92.0%	96.0%

Table 3: Cumulative parsing success rates across models and processing stages ($N = 880$; representing 110 documents across 4 tasks and 2 text modes). R1: after initial run; +R: after reruns (up to 2); +Rep: after ad-hoc regex repair parsing.

Task Type	Initial	Final
T1: Extraction	91.5%	100%
T2: Isolated Norm	80.1%	94.0%
T3: Pipeline Norm	86.8%	97.0%
T4: Combined Extr/Norm	75.4%	92.0%

Table 4: Parsing success rates per task, both initially and after reruns and ad-hoc parsing.

Finally, outputs from all runs were merged using a quality-priority system that prioritised valid parsed data over salvaged JSON. Documents for which no parseable output could be obtained after these steps were treated as empty outputs and assigned a score of 0.0 for all metrics to penalise model fragility.

The choice of JSON as the output format was intended to facilitate the direct integration of extracted data into downstream pipelines and to ensure clear pairing between historical and normalised phrases in combined tasks. While simpler formats (e.g., plain-text lists) might have reduced syntax errors, they might also have increased the risk of ambiguity when mapping multiple entities. Our evaluation of structural reliability (Table 3) shows that while GPT-4o and Llama 3.3 70B achieved near-perfect JSON adherence, the smaller models Llama 3.1 8B and Mixtral 8x22B were more fragile. Our recovery protocol, including reruns and ad-hoc regex repair parsing, significantly improved final parsing success, particularly for Llama 8B.

In addition, we find that rhetorical filtering has a negligible effect on the structural success of the output, where the average success rate was 83.1% averaged for all models after the first run and 98.9% after post-hoc parsing, compared to the full text version of 79.2% and 98.3%, respectively. The nature of the NLP task and requested output, however, is a significant predictor of model fragility, as can be seen in Table 4. While the more simple and straightforward extraction (Task 1) reached 91.5% initial success rate, the more complex combined extraction and normalisation task (Task 4) yielded only 75.4%. This likely reflects the increased dif-

ficulty of maintaining nested JSON syntax during simultaneous linguistic transformations.

4.4. Evaluation

To evaluate the correspondence between model-predicted and gold-standard work phrases, we apply a combination of **exact-match**, **string-level**, and **embedding-based** similarity metrics. This multi-level evaluation captures both surface-form correspondence and semantic adequacy.

Before evaluation, all phrases are lowercased and tokenised. Metrics are computed on a per-document basis and aggregated across the corpus for each prompt variant, enabling comparison of model performance under different input conditions (full text vs. rhetorically filtered) and task configurations.

Duplicate handling: Since the same work phrase may appear multiple times within a document, we preserve all instances during evaluation. For exact match metrics, if a phrase appears n times in gold and m times in predictions, we count $\min(n, m)$ as true positives. For alignment-based metrics (Levenshtein and embedding similarity), each phrase instance is matched independently, ensuring models are credited for extracting all occurrences of repeated phrases.

4.4.1. Exact Match Metrics

We compute **precision**, **recall**, and **F1-score** based on *exact string identity* at both phrase and token levels, counting true positives as $\sum \min(\text{count}_{\text{gold}}(p), \text{count}_{\text{pred}}(p))$ across unique phrase types p. Token-level metrics apply the same principle to individual tokens. Together, these provide complementary insights into fully versus partially correct outputs.

4.4.2. String-Level Similarity

To capture surface-level similarity beyond exact matches, we compute *normalised Levenshtein similarity* between gold and predicted phrases, allowing partial matches to be recognised despite small spelling or formatting differences. Levenshtein distance counts the minimum number of insertions, deletions, or substitutions to transform one string into another, converted to a similarity score in $[0, 1]$:

$$\text{Sim}_{\text{lev}}(g_i, p_j) = 1 - \frac{d(g_i, p_j)}{\max(|g_i|, |p_j|)},$$

where $d(g_i, p_j)$ is the character edit distance. A score of 1 indicates identical strings, 0 indicates maximal difference.

Since the number of predicted phrases may differ from the gold standard, we treat alignment as a matching problem and compute two complementary scores per document:

- **Gold-based (recall-oriented):** measures how well each gold phrase is recovered,

$$\text{LevSim}_{G \to P} = \frac{1}{|G|} \sum_{g_i \in G} \max_{p_j \in P} \text{Sim}_{\text{lev}}(g_i, p_j),$$

- **Prediction-based (precision-oriented):** measures how well each predicted phrase matches a gold phrase,

$$\text{LevSim}_{P \to G} = \frac{1}{|P|} \sum_{p_j \in P} \max_{g_i \in G} \text{Sim}_{\text{lev}}(g_i, p_j).$$

Phrases without a matching counterpart are assigned a similarity of 0.

4.4.3. Embedding-Based Semantic Similarity

To capture semantic similarity beyond surface form, we compute the cosine similarity between mean phrase embeddings, using optimal phrase-to-phrase alignment, allowing recognition of semantically equivalent phrases despite differing wording. Historical phrases use a domain-specific fastText model trained on diachronic Swedish text (Hengchen and Tahmasebi, 2021), specifically the incrementally trained vectors spanning the period 1740–1800 to best match our data. While transformer models are state-of-the-art for many tasks, we use static embeddings because our data consists of isolated phrases lacking sufficient context. A historical fastText model further helps align with older language and improves robustness to non-standard spelling by representing words through character-level n-grams. We here match the model's training constraints by removing non-Swedish characters and tokens of length two or smaller. Normalised phrases use a Swedish fastText model trained on Wikipedia (Grave et al., 2018). Both models use subword information for robust out-of-vocabulary handling. However, should there be phrases where no valid tokens remain after preprocessing, we assign a zero vector, yielding a similarity of 0 when compared to any phrase. To obtain a single embedding per phrase, we compute the average of all word embeddings within each phrase.

Given gold phrases G and predicted phrases P, we construct a similarity matrix S where S_{ij} is the cosine similarity between embeddings of gold phrase g_i and predicted phrase p_j:

$$S_{ij} = \frac{v(g_i) \cdot v(p_j)}{\|v(g_i)\| \, \|v(p_j)\|},$$

where -1 indicates maximally opposed meaning, 0 no semantic relationship (orthogonal vectors), and 1 perfect alignment.

As with string-level metrics, predicted and gold phrases may differ in number or order. We perform optimal phrase-to-phrase matching, assigning unmatched phrases a similarity of 0. We report two complementary perspectives analogous to the Levenshtein alignment:

- **Gold-based (recall-oriented):**

$$\text{Sim}_{G \to P} = \frac{1}{|G|} \sum_{g_i \in G} \max_{p_j \in P} S_{ij},$$

  measuring how well predicted phrases cover gold phrases.

- **Prediction-based (precision-oriented):**

$$\text{Sim}_{P \to G} = \frac{1}{|P|} \sum_{p_j \in P} \max_{g_i \in G} S_{ij},$$

  measuring how well predicted phrases correspond to gold references.

5. Results and Discussion

We evaluate four LLMs across four work-phrase extraction and normalisation tasks. Results are reported for both full original texts and filtered versions (with *Salutatio* and *Conclusio* removed). All results are reported as means ± standard deviation.

5.1. Overall Performance and Output Statistics

Identifying phrases describing work in historical petitions remains a significant challenge. As shown in Table 5, a primary driver of low precision is the tendency of LLMs to over-produce content. While the gold standard averages 4.81 phrases per document, models, particularly Llama 8B, consistently over-generate candidates (13.65 phrases). This high output volume directly explains why recall consistently exceeds precision: by over-generating candidates, models increase the likelihood of overlapping with gold spans at a high cost to precision. Furthermore, the extracted historical phrases are consistently shorter than those in the gold standard. This suggests that models sometimes fail to capture the full extent of annotated work-related spans, or that responses are being prematurely truncated during JSON generation. However, this length discrepancy does not extend to the normalisation process, since the predicted normalised phrases remain close to the gold standard in length.

To provide a reference for the embedding-based metrics, we establish a domain-internal baseline by calculating the mean similarity between randomly paired gold and predicted phrases across the test set. This provides a baseline score of 0.72 for historical phrases (used in Tasks 1, 3, and 4) and 0.51 for normalized phrases (used in Tasks 2, 3, and 4). The higher historical baseline reflects the formulaic nature and lexical density of 18th-century petitionary language, where work-related expressions frequently share domain-specific semantics. Consequently, embedding scores in the following sec-

Source	Mode	Phrases/Doc	Tokens/Phrase
Gold Hist.	-	4.81 ± 4.50	6.32 ± 4.62
Gold Norm.	-	4.81 ± 4.50	2.80 ± 0.87
Task 1: Phrase Extraction			
Llama 70B	Full	7.85 ± 5.88	3.97 ± 1.87
	Filt.	7.81 ± 5.90	4.08 ± 1.96
Llama 8B	Full	13.65 ± 9.58	4.98 ± 2.56
	Filt.	13.64 ± 9.59	4.99 ± 2.65
Mixtral	Full	5.85 ± 4.10	4.42 ± 2.16
	Filt.	6.44 ± 5.83	4.81 ± 3.12
GPT-4o	Full	6.85 ± 5.06	5.66 ± 2.20
	Filt.	6.72 ± 5.54	5.58 ± 2.17
Task 2: Isolated Normalisation (Gold Input)			
Llama 70B	Full	13.35 ± 10.31	2.69 ± 0.60
	Filt.	11.10 ± 5.87	2.71 ± 0.67
Llama 8B	Full	11.64 ± 8.55	2.95 ± 1.28
	Filt.	13.23 ± 14.20	3.21 ± 2.01
Mixtral	Full	8.26 ± 5.17	2.27 ± 0.59
	Filt.	8.76 ± 5.69	2.24 ± 0.50
GPT-4o	Full	7.04 ± 5.54	2.63 ± 0.81
	Filt.	6.35 ± 5.12	2.53 ± 0.61

Table 5: Descriptive Statistics (Mean ± SD) for Tasks 1 and 2. Predicted outputs are shown for both original full text and filtered text modes.

Metric	Mode	L70B	L8B	Mixtral	GPT
Phrase P	Full	.069±.15	.041±.13	**.099±.20**	.077±.17
	Filt.	**.098±.19**	.042±.13	.081±.17	.077±.17
Phrase R	Full	.097±.21	.089±.20	.102±.20	**.147±.30**
	Filt.	.116±.22	.090±.20	.086±.18	**.147±.30**
Phrase F1	Full	.071±.15	.049±.13	**.087±.16**	.084±.16
	Filt.	**.098±.18**	.049±.13	.073±.14	.084±.16
Token P	Full	.272±.24	.196±.19	.298±.23	**.301±.21**
	Filt.	.303±.24	.197±.19	.270±.21	**.310±.20**
Token R	Full	.364±.28	.471±.27	.351±.28	**.513±.30**
	Filt.	.393±.27	.473±.27	.349±.28	**.495±.27**
Token F1	Full	.255±.18	.224±.16	.279±.20	**.319±.18**
	Filt.	.294±.20	.227±.16	.253±.17	**.329±.16**
Lev (G)	Full	.378±.20	.398±.19	.354±.23	**.427±.26**
	Filt.	.397±.20	.397±.19	.357±.22	**.427±.26**
Lev (P)	Full	.251±.19	.168±.16	**.293±.22**	.280±.18
	Filt.	.278±.21	.168±.16	.277±.19	**.280±.18**
Emb (G)	Full	.784±.16	**.836±.13**	.707±.23	.781±.21
	Filt.	.796±.15	**.833±.13**	.722±.23	.781±.21
Emb (P)	Full	.518±.28	.354±.26	.567±.30	**.568±.28**
	Filt.	.528±.28	.354±.26	.561±.28	**.568±.28**

Table 6: Task 1 Extraction Performance: Mean ± SD for Precision (P), Recall (R), and F1-score (F1). (G) denotes Gold-oriented (recall-like) and (P) denotes Prediction-oriented (precision-like) similarities. Comparison of Full original and Filtered (Filt.) text modes across all models.

Metric	Mode	L70B	L8B	Mixtral	GPT
Task 2: Isolated Normalisation					
Phrase P	Full	.023±.06	.011±.04	.037±.12	**.077±.19**
	Filt.	.032±.09	.009±.04	.059±.18	**.085±.21**
Phrase R	Full	.050±.14	.023±.11	.039±.10	**.088±.19**
	Filt.	.047±.11	.012±.05	.066±.17	**.089±.19**
Phrase F1	Full	.027±.06	.013±.04	.033±.09	**.073±.16**
	Filt.	.035±.08	.010±.04	.049±.13	**.077±.16**
Token F1	Full	.124±.10	.096±.08	.121±.13	**.196±.16**
	Filt.	.139±.12	.091±.08	.142±.16	**.207±.16**
Lev (G)	Full	**.446±.14**	.374±.14	.342±.20	.414±.18
	Filt.	**.437±.14**	.375±.13	.364±.22	.414±.20
Emb (G)	Full	**.716±.11**	.649±.15	.559±.27	.650±.19
	Filt.	**.704±.11**	.651±.14	.575±.26	.628±.21
Task 3: Pipeline Normalisation					
Phrase P	Full	.049±.10	.026±.11	.060±.15	**.089±.19**
	Filt.	.055±.13	.025±.11	.058±.14	**.089±.18**
Phrase R	Full	.069±.15	.033±.12	.066±.15	**.082±.16**
	Filt.	.081±.18	.033±.12	.063±.13	**.086±.19**
Phrase F1	Full	.053±.10	.024±.08	.055±.12	**.080±.16**
	Filt.	.059±.12	.024±.08	.055±.11	**.081±.17**
Token F1	Full	.157±.13	.095±.10	.166±.15	**.217±.17**
	Filt.	.164±.13	.093±.10	.176±.15	**.217±.16**
Lev (G)	Full	**.447±.14**	.376±.15	.373±.19	.413±.18
	Filt.	**.447±.15**	.371±.16	.387±.18	.408±.18
Emb (G)	Full	**.705±.12**	.640±.15	.599±.23	**.631±.20**
	Filt.	**.703±.12**	.631±.16	.618±.20	.626±.20

Table 7: Task 2 and Task 3 Normalisation Performance: Mean ± SD for Precision (P), Recall (R), and F1-score (F1). (G) denotes Gold-oriented (recall-like) and (P) denotes Prediction-oriented (precision-like) similarities. Comparison of Full original and Filtered (Filt.) text modes across all models.

Metric	Mode	L70B	L8B	Mixtral	GPT
Combined Task: Extraction Component					
Phrase P	Full	.083±.17	.033±.08	.081±.19	**.086±.16**
	Filt.	.081±.17	.033±.08	.067±.13	**.088±.14**
Phrase R	Full	.100±.22	.062±.15	.090±.19	**.141±.28**
	Filt.	.103±.22	.062±.15	.081±.17	**.149±.28**
Phrase F1	Full	.083±.17	.038±.09	.078±.16	**.094±.17**
	Filt.	.081±.17	.038±.09	.067±.12	**.098±.16**
Token F1	Full	.278±.21	.216±.15	.273±.18	**.316±.19**
	Filt.	.294±.20	.216±.15	.253±.17	**.311±.19**
Lev (G)	Full	.377±.21	.366±.17	.350±.21	**.411±.24**
	Filt.	.381±.21	.366±.17	.345±.21	**.398±.25**
Emb (G)	Full	.765±.18	**.813±.16**	.734±.21	.776±.20
	Filt.	.768±.18	**.813±.16**	.716±.24	.755±.23
Combined Task: Normalisation Component					
Phrase F1	Full	.014±.05	.010±.04	.026±.07	**.027±.09**
	Filt.	.020±.07	.010±.04	.021±.06	**.032±.09**
Token F1	Full	.141±.12	.087±.09	.131±.12	**.155±.14**
	Filt.	.146±.12	.087±.09	.129±.12	**.168±.14**
Lev (G)	Full	.338±.14	.332±.15	.325±.15	**.340±.15**
	Filt.	.346±.14	.332±.15	.323±.15	**.349±.17**
Emb (G)	Full	.623±.17	.630±.18	.571±.20	**.631±.18**
	Filt.	**.629±.16**	**.629±.18**	.570±.20	.624±.20

Table 8: Task 4 Combined Extraction and Normalisation Performance: Mean ± SD for Precision (P), Recall (R), and F1-score (F1). (G) denotes Gold-oriented (recall-like) and (P) denotes Prediction-oriented (precision-like) similarities. Comparison of Full original and Filtered (Filt.) text modes across all models.

tions must be interpreted relative to these thresholds.

5.2. Task 1: Extraction of Phrases Describing Work

Task 1 measures the models' ability to identify historical work-phrase spans. As shown in Table 6, models generally struggle with exact span reproduction. However, token-level F1 consistently exceeds phrase-level F1 by a factor of three, and Embedding Similarity scores average ≈ 0.80. While these scores indicate that models consistently locate relevant topical regions, they are only slightly above the 0.72 random baseline, suggesting that while the general semantic content is captured, the models struggle to isolate the specific nuances or precise boundaries defined by human annotators.

5.3. Normalisation: In Isolation vs. Pipeline

Here, we evaluate the models' ability to normalise historical phrases describing work into modern, standardised Swedish. Tasks 2 and 3 compare the linguistic capability of normalisation in isolation (based on manually extracted gold phrases) against a more demanding pipeline setting (based on phrases extracted by the system). Descriptive statistics for these tasks are summarised in Table 7.

Interestingly, Task 2 (Isolated) proved more difficult than Task 3 (Pipeline) for all models. Exact phrase matching remained low ($F1 < .09$), but performance improved when moving from isolated phrases to the pipeline. This suggests a benefit from surrounding context cues: performing extraction and normalisation in sequence allows models to leverage broader semantic cues to resolve historical ambiguities that are absent when phrases are processed in isolation.

5.4. Task 4: Combined Extraction and Normalisation

Task 4 tests whether a joint approach functions as an implicit chain-of-thought. As shown in Table 8, for GPT, this combined task acted as a helpful guide; its extraction score improved from .084 (Task 1) to .098 (Task 4). This suggests that, for very strong models, the extra effort of normalising a phrase may help them "sharpen" their focus and identify the correct historical span.

In contrast, other models seem overwhelmed by the dual-task requirement. For Llama 70B, doing both tasks at once appears to be a distraction, causing its extraction score to drop from .098 to .081. Most importantly, the quality of the normalisation collapse for all models in this combined setup

compared to the staged pipeline. For instance, GPT's normalisation score falls from .081 in Task 3 to only .032 in Task 4. These results indicate that while combined processing may sharpen initial span detection, it compromises the lexical precision required for accurate normalisation.

5.5. Cross-Task Comparisons and Analysis of Text Filtering

A comparison of all four tasks reveals a distinct difficulty hierarchy: Task 4 (Normalisation Component) is the most difficult, followed by Task 2 (Isolated Normalisation). Paradoxically, models perform best on the Extraction Component of Task 4. This suggests that while multitasking helps models locate relevant text, the complexity of producing a dual-language output somewhat degrades their ability to perform the linguistically demanding task of normalisation.

The impact of text filtering (removing *Salutatio* and *Conclusio*) is generally marginal and inconsistent. Across most tasks, the difference between Full and Filtered modes is minor. Furthermore, the effect is not uniform across models; while filtering provides a small boost to normalisation in the stronger models (GPT and L70B) by reducing noise, it occasionally hinders extraction for others. For example, Mixtral's extraction performance actually drops in Filtered mode in both Task 1 and Task 4. This suggests that while noise reduction can slightly help with the linguistic focus required for normalisation, the models are largely robust to the structural "noise" of historical petitions, making filtering a secondary factor in overall performance.

5.6. Qualitative Error Analysis

To complement the quantitative evaluation, we conduct a targeted qualitative error analysis. We manually examine cases in which the models perform poorly, as well as targeting instances where surface-level and semantic-level evaluation metrics diverge.

A recurrent issue is the failure to produce parseable output, with Llama 8B and Mixtral exhibiting the highest failure rates (see more in Section 4.3). Llama 8B most often produces malformed outputs, likely reflecting the limitations of a smaller model handling complex formatting instructions. By contrast, Mixtral frequently returns empty outputs or echoes the input, suggesting that technical factors like API instability or system load contribute to these failures. Notably, occasional parse failures occur for all models without an apparent structural cause, illustrating the unpredictability of LLM behavior.

Another prominent error pattern in low-performing cases is systematic over-prediction. Models frequently identify more phrases than present in the gold data, consistent with higher recall than precision. In extreme cases, several models tag a wide range of isolated verbs as work-related, such as "emottaga" (*receive*) and "frågat" (*asked*), even when the surrounding context does not support such interpretations. These errors are not limited to verbs, but also instances of nouns. Longer input texts appear to amplify this tendency toward overgeneralisation. We further observe sporadic hallucinations, i.e., predictions that cannot be traced to any span in the input text. However, a larger-scale analysis would be needed to determine their frequency.

At the same time, we identify cases in which models predict phrases that are not present in the gold annotations but can plausibly describe work-related activities when considered in context. Examples include verb phrases such as "brukat Qwacksalverÿ och ögonsalvor" (*practised quackery and eye-ointments*), which are normalised to "bruka kvacksalveri" (*practice quackery*) and "tillverka/sälja/bruka ögonsalvor" (*make/sell/use eye salves*), as well as the normalised phrase "tjäna som fältskärs gesäll" (*serve as an apprentice/journeyman barber-surgeon*). There are also several cases of occupational titles being extracted, including "Rådman" (*councilman*) and "Inspector" (*inspector*), which are normalised to "arbeta/tjäna som rådman" (*work as a councilman*), "arbeta/tjäna som inspektor" (*work as an inspector*), or "utföra inspektion" (*perform inspection*). Consultation with historians suggests that, although models often overgeneralise, some outputs seem to highlight valid work activities that could enrich manual curation.

Comparisons between surface-level (Levenshtein) and semantic (embedding-based) similarity reveal frequent divergences. Low string similarity often arises when predicted spans overlap with gold spans but include additional material, or when character overlap is incidental. Because each metric aligns predictions independently, direct one-to-one interpretation is somewhat limited. Moreover, the metrics capture fundamentally different aspects of similarity, which further limits direct one-to-one interpretation. Nevertheless, some consistent patterns emerge. When a prediction preserves most of the intended meaning but diverges in surface form, embedding similarity tends to align more closely with human judgement. For example:

Gold: "stå uthe på fiskaretorget"
Prediction: "stå uthe på fiskaretorget att förtiäna något till uppehälle"
Gold translated: *stand outside in the fishermen's square*
Prediction translated: *stand outside in the fishermen's square to earn something for a living*
Levenshtein score: 0.43
Embedding score: 0.93

However, embedding similarity also exhibits behaviour that appears counterintuitive from a hu-

man perspective. High scores do not always correspond to strong semantic equivalence, and conversely, low scores can occur even when the core work-related activity is preserved. This variability indicates that embedding-based similarity can diverge from human judgement in both directions, making it difficult to interpret in isolation. As established in Section 4.1, embedding scores must be interpreted in light of the high domain-internal baselines (0.72 for historical text and 0.51 for normalised text). A score slightly above these thresholds risks overstating semantic alignment, while scores below the baseline may under-represent semantic proximity between near-synonyms. Examples from both Task 1 (the extraction task, with historical original phrases) and Task 3 (pipeline normalisation, with normalised modern phrases) include:

Gold: "sälja varor"
Prediction: "bedriva handel"
Gold translated: *sell goods*
Prediction translated: *conduct trade*
Embedding score: 0.66

Gold: "bruka stenbod"
Prediction: "bruka lokal"
Gold translated: *use a stone shop*
Prediction translated: *use premises*
Embedding score: 0.78

Gold: "taget penningar till låns"
Prediction: ""smida för sin räcḧning"'
Gold translated: *borrowed money*
Prediction translated: *forge for one's own account*
Embedding score: 0.80

Taken together, these observations suggest that embedding-based similarity provides a valuable signal for identifying partially correct extractions, but also carries a risk of overstating semantic alignment between gold annotations and model outputs. A combined evaluation strategy that integrates both surface-level and semantic metrics therefore appears necessary for a more comprehensive and reliable assessment of model behaviour than relying on either metric alone.

6. Conclusions

This study explores the potential of four large language models, GPT-4o, Llama-3 (70B and 8B), and Mixtral-8x7B, for extracting and normalising phrases describing work from historical Swedish petitions. Our experiments demonstrate that while LLMs struggle to reproduce human-annotated spans exactly, yielding low phrase-level precision and F1 scores, token-level and embedding-based metrics offer complementary perspectives. These results indicate that models consistently locate relevant topical regions, although the high domain-specific baselines suggest that embedding scores can reflect general semantic intent rather than precise equivalence.

The choice of task architecture emerges as a factor in performance. We find that normalisation is notably more effective when performed within a contextual pipeline (Task 3) than in isolation (Task 2), suggesting that historical normalisation relies on integrative semantic cues. Furthermore, our results reveal a "multitasking paradox" in combined setups: while simultaneous extraction and normalisation can sharpen phrase identification for high-parameter models, the increased complexity significantly degrades the lexical precision of the normalised output. Although rhetorical filtering may prove valuable for other downstream applications, where removing formulaic sections could more effectively improve signal quality, the removal of *Salutatio* and *Conclusio* sections show only marginal effects in the present tasks.

Manual error analysis highlights a systematic tendency toward over-prediction and occasional hallucinations. However, the models also identify several plausible work-related expressions absent from the human annotations, pointing to the potential value of LLMs for discovery-driven research. A more extensive manual evaluation of these candidates would be valuable for understanding how much genuinely new information can be recovered versus the amount of noise added through overgeneralisation.

Overall, our results indicate that LLMs show promise for facilitating the identification of work-related activities in historical texts, provided their outputs are subject to expert evaluation and interpreted through metrics that account for high background similarity. Future work may explore more robust prompting strategies, domain-adapted fine-tuning, and hybrid human–machine workflows that combine automated semantic detection with expert validation. While a staged pipeline currently offers the most reliable balance of accuracy, the models' ability to detect valid expressions missed by humans suggests that such workflows could significantly enhance the coverage and interpretability of cultural heritage datasets.

7. Limitations

While this study provides insights into the capabilities of LLMs for processing historical Swedish, several limitations should be noted.

First, our evaluation is limited to a specific genre of historical text, 18th-century petitions, and a single language. The linguistic challenges found here, such as non-standardised orthography and formulaic rhetorical structures, may behave differently in other historical contexts, time periods, or languages.

Second, the gold standard used for evaluation, while expert-annotated, represents only one possible interpretation of phrases describing work. As noted in our qualitative analysis, the models sometimes identify potential work activities that are not present in the gold data, suggesting that exact-match metrics like F1-score may provide an overly conservative estimate of model utility in discovery-driven research.

Third, while we made use of twelve purposefully selected, diverse examples to ground the models across a range of linguistic challenges, performance in LLMs can be highly sensitive to prompt wording and the specific selection and ordering of examples. Although our prompt template was iteratively refined, we did not perform an thorough search of the prompt space; thus, our instructions may represent a suboptimal configuration for certain model architectures. Furthermore, the instructional complexity and total prompt length required to convey these nuanced historical tasks likely strained the context-handling capabilities of smaller models, contributing to the observed output instability and malformed JSON responses.

Finally, while few-shot prompting provides a strong baseline, this study did not explore domain-specific fine-tuning or more advanced reasoning architectures, which could improve the balance between extraction and normalisation accuracy in the combined task setup.

8. Ethical Considerations

The use of large language models in historical research raises several ethical considerations. Although large models entail energy costs, our few-shot prompting approach avoids resource-intensive fine-tuning and thus has a comparatively lower environmental impact. The historical petitions used in this study are in the public domain and contain no modern personal data, and their use follows established ethical guidelines for historical sources. Finally, our approach is designed to support, not replace, human expertise. We adopt a human-in-the-loop perspective, where automated extraction serves to assist and enrich manual curation rather than provide definitive interpretations, helping to guard against the uncritical acceptance of model-generated errors as historical fact.

9. Acknowledgments

We are grateful to our colleagues in the Petitions and GaW projects: Jonas Lindström, Örjan Kardell, Jezzica Israelsson, Maria Ågren, Sofia Ling, Linda Oja, and Fredrik Wahlberg, for their insightful discussions and collaborative support. Finally, we appreciate the constructive feedback from the anonymous reviewers. This research was funded by the Swedish Research Council (grant number 2018-06159).

10. Bibliographical References

Maria Ågren, Jonas Lindström, and Sofia Ling. 2023. The principles of entering data into the Gender and Work database. Technical report, Uppsala University and Umeå University. Confirmed by the management of the GaW database 2023-12-11.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. *arXiv preprint arXiv:1904.02036*.

Toby Burrows. 2023. Reproducibility, verifiability, and computational historical research. *Journal on Computing and Cultural Heritage*, 16(3):1–12.

Kevin Cohen, Laura Manrique-Gómez, and Rubén Manrique. 2025. Historical ink: Exploring large language models for irony detection in 19th-century Spanish. *arXiv preprint arXiv:2503.22585*.

Rosemarie Fiebranz, Erik Lindberg, Jonas Lindström, and Maria Ågren. 2011. Making verbs count: the research project 'gender and work' and its methodology. *Scandinavian Economic History Review*, 59(3):273–293.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Stina Hansson. 1988. *Svensk brevskrivning: teori och tillämpning*, volume 18. Göteborgs universitet.

Simon Hengchen and Nina Tahmasebi. 2021. A collection of Swedish diachronic word embedding models trained on historical newspaper data. *Journal of Open Humanities Data*, 7(2):1–7.

Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, et al. 2025. Ner4all or context is all you need: Using llms for low-effort, high-performance ner on historical texts. a humanities informed approach. *arXiv preprint arXiv:2502.04351*.

Mark Humphries, Lianne C Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella

Murray, and Elizabeth Spence. 2025. Unlocking the archives: Using large language models to transcribe handwritten historical documents. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pages 1–19.

Jezzica Israelsson. 2016. In consideration of my meagre circumstances: The language of poverty as a tool for ordinary people in early modern Sweden. Master's thesis, Uppsala Universitet.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Lucija Krusic. 2024. Constructing a sentiment-annotated corpus of austrian historical newspapers: Challenges, tools, and annotator experience. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 51–62.

Ellinor Lindqvist, Eva Pettersson, and Joakim Nivre. 2025. Finding the plea: Evaluating the ability of llms to identify rhetorical structure in swedish and english historical petitions. In *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*, pages 86–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

OpenAI. 2026. GPT-4o. API model accessed January 2026.

Michael Piotrowski. 2012. *Natural language processing for historical texts*, 1 edition. Synthesis lectures on human language technologies. Morgan Claypool Publishers, Cham.

Thomas Sokoll. 2006. Writing for relief: Rhetoric in English pauper letters, 1800–1834. In Andreas Gestrich, Steven King, and Lutz Raphael, editors, *Being Poor in Modern Europe: Historical Perspectives 1800–1940*, pages 91–112. Peter Lang, Bern.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Crina Tudor, Beáta Megyesi, and Robert Östling. 2025. Prompting the past: Exploring zero-shot learning for named entity recognition in historical texts using prompt-answering llms. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 216–226.

Huub Wijfjes. 2017. Digital humanities and media history: A challenge for historical newspaper research. *Digital Journalism*, 5(1):29–43.

David C. S. Wilson. 2022. Working at scale: What do computational methods mean for research using collections, models and big data? *Science Museum Group Journal*, (18).

Appendix: Prompts Used in Experiments

Prompt Task 1: Extraction

Swedish Original

```
Du är expert på lingvistisk analys av svenska historiska texter. Din uppgift är att
    analysera svenska suppliker från 1700-talet och identifiera alla fraser som
    beskriver arbete, arbetsutförande eller arbetsrelaterade handlingar.

# Huvudinstruktioner
1. **Läs texten noggrant.** Texten kan innehålla gammaldags stavning, fri ordföljd
    och sammanskrivningar.
2. **Extrahera endast fraser som beskriver arbete, arbetsutförande eller
    arbetsrelaterade handlingar.** Detta inkluderar att utföra, bidra till eller
    delta i arbete, syssla, tjänst eller plikt för försörjning, oavsett om ersä
    ttning ges.
3. **Bevara originaltexten exakt.** Lägg INTE till, ta bort eller ändra ord,
    stavning eller skiljetecken.
4. **Returnera en lista över extraherade fraser** i den ordning de förekommer i
    texten.

# Outputformat
Returnera endast ett JSON-objekt med nyckeln **"extracted_phrases"** och en lista
    med fraser från texten. Returnera JSON exakt i detta format, utan extra text,
    kommentarer, sammanfattningar, extra nycklar eller kodblock-markeringar:

{
  "extracted_phrases": ["mala någon säd", "i tiänst warit"]
}

# Exempel
(Endast korta utdrag visas i exemplen. I uppgiften ska hela texten bearbetas.)

### Input 1:
"/.../ efter som thenne tiden warit så swårt efter humble? Swarades hon hade kiöpt
    humbla ifrån Tidön, hwartil hon giordt en liten tilökning af några Malörts
    Qwistar, /.../"

### Output 1:
{
  "extracted_phrases": ["kiöpt humbla"]
}

/.../

# Identifiera och extrahera de fraser som beskriver arbete i följande supplik:
```

English Translation

```
You are an expert in the linguistic analysis of Swedish historical texts. Your task
    is to analyse Swedish petitions from the 18th century and identify all phrases
    describing work, the performance of work, or work-related actions.

# Main Instructions
1. **Read the text carefully.** The text may contain archaic spelling, free word
    order, and compound words.
2. **Extract only phrases describing work, the performance of work, or work-related
    actions.** This includes performing, contributing to, or participating in work,
    occupation, service, or duty for a livelihood, regardless of whether
    compensation is provided.
3. **Preserve the original text exactly** Do NOT add, remove, or change words,
    spellings, or punctuation marks.
4. **Return a list of extracted phrases** in the order they appear in the text.

# Output Format:
Return only a JSON object with the key **"extracted_phrases"** and a list of phrases
     from the text. Return the JSON exactly in this format, without any additional
    text, comments, summaries, extra keys, or code block markers:

{
  "extracted_phrases": ["grind some grain", "been in service"]
}

# Examples
(Only short excerpts are shown in the examples. In the actual task, the entire text
    should be processed.)

### Input 1:
"/.../ seeing as hops have been so scarce of late? She answered that she had
    purchased hops from Tidö, to which she had made a small addition of some
    Wormwood Twigs, /.../"

### Output 1:
{
  "extracted_phrases": ["purchased hops"]
}

/.../

# Identify and extract the phrases describing work in the following petition:
```

Prompt Task 2: Normalisation in Isolation

Swedish Original

```
Du är expert på lingvistisk analys av svenska historiska texter. Din uppgift är att
    normalisera fraser som beskriver arbete, arbetsutförande eller arbetsrelaterade
    handlingar i svenska suppliker från 1700-talet. Du kommer att få en lista med
    extraherade fraser som har identifierats som beskrivningar av arbete. Du ska
    endast normalisera dessa fraser.

# Huvudinstruktioner
1. **Läs varje fras noggrant.** Fraserna kan innehålla gammaldags stavning,
    grammatik och ordföljd.
2. **Normalisera varje fras till modern svenska i grundform (infinitiv).**
   - Använd modern stavning och modern grammatik.
   - Formulera varje fras som en verbfras i infinitiv, även om originalfrasen inte
     är verbal.
3. **Ändra inte innehållet.**
   Lägg inte till, ta bort eller slå ihop fraser.
4. **Bevara ordningen.** Varje normaliserad fras ska motsvara frasen på samma
    position i indata.

# Outputformat
Returnera endast ett JSON-objekt med nyckeln **"normalised_phrases"** och en lista
    med fraser i normaliserad grundform. Returnera JSON exakt i detta format, utan
    extra text, kommentarer, sammanfattningar, extra nycklar eller kodblock-
    markeringar:

{
  "normalised_phrases": ["mala säd", "vara i tjänst"]
}

# Exempel

### Input 1:
{
  "extracted_phrases": ["kiöpt humbla"]
}

### Output 1:
{
  "normalised_phrases": ["köpa humle"]
}

# Normalisera följande fraser:
```

English Translation

```
You are an expert in the linguistic analysis of Swedish historical texts. Your task
    is to analyse Swedish petitions from the 18th century and identify all phrases
    describing work, the performance of work, or work-related actions.

You are an expert in the linguistic analysis of Swedish historical texts. Your task
    is to normalise phrases describing work, the performance of work, or work-
    related actions in Swedish petitions from the 18th century. You will be provided
     with a list of extracted phrases that have been identified as descriptions of
    work. You are to only normalise these phrases.

# Main Instructions
1. **Read each phrase carefully.** The phrases may contain archaic spelling, grammar
    , and word order.
2. **Normalise each phrase into modern Swedish in the base form (infinitive)**
   - Use modern spelling and modern grammar.
   - Formulate each phrase as an infinitive verb phrase, even if the original phrase
       is not verbal.
3. **Do not change the content.**
   Do not add, remove, or merge phrases.
4. **Preserve the order.** Each normalised phrase must correspond to the phrase at
    the same position in the input data.

# Output Format:
Return only a JSON object with the key **"normalised_phrases"** and a list of
    phrases in normalised base form. Return the JSON exactly in this format, without
     any additional text, comments, summaries, extra keys, or code block markers:

{
  "normalised_phrases": ["grind grain", "to be in service"]
}

# Examples

### Input 1:
{
  "extracted_phrases": ["purchased hops"]
}

### Output 1:
{
  "normalised_phrases": ["purchase hops"]
}

/.../

# Normalise the following phrases::
```

Prompt Task 3: Normalisation Pipeline

Swedish Original

```
Du är expert på lingvistisk analys av svenska historiska texter. Din uppgift är att
    analysera svenska suppliker från 1700-talet och identifiera alla fraser som
    beskriver arbete, arbetsutförande eller arbetsrelaterade handlingar, och återge
    dessa i normaliserad grundform (infinitiv).

# Huvudinstruktioner
1. **Läs texten noggrant.** Texten kan innehålla gammaldags stavning, fri ordföljd
    och sammanskrivningar.
2. **Identifiera endast fraser som beskriver arbete, arbetsutförande eller
    arbetsrelaterade handlingar.** Detta inkluderar att utföra, bidra till eller
    delta i arbete, syssla, tjänst eller plikt för försörjning, oavsett om ersä
    ttning ges.
3. **Normalisera varje identifierad fras till modern grundform.**
    Använd modern stavning och grammatik.
    Formulera varje fras som en verbfras i infinitiv, även om originalet inte är
       verbalt.
4. **Returnera en lista över normaliserade fraser** i den ordning de förekommer i
    texten.

# Outputformat
Returnera endast ett JSON-objekt med nyckeln **"normalised_phrases"** och en lista
    med fraser i normaliserad grundform. Returnera JSON exakt i detta format, utan
    extra text, kommentarer, sammanfattningar, extra nycklar eller kodblock-
    markeringar:

{
  "normalised_phrases": ["mala säd", "vara i tjänst"]
}

# Exempel
(Endast korta utdrag visas i exemplen. I uppgiften ska hela texten bearbetas.)

### Input 1:
"/.../ efter som thenne tiden warit så swårt efter humble? Swarades hon hade kiöpt
    humbla ifrån Tidön, hwartil hon giordt en liten tilökning af några Malörts
    Qwistar, /.../"

### Output 1:
{
  "normalised_phrases": ["köpa humle"]
}

# Identifiera och normalisera (i grundform) de fraser som beskriver arbete i
  följande supplik:
```

English Translation

```
You are an expert in the linguistic analysis of Swedish historical texts. Your task
    is to analyse Swedish petitions from the 18th century and identify all phrases
    describing work, the performance of work, or work-related actions, and provide
    these in normalised base form (infinitive).

# Main Instructions
1. **Read the text carefully.** The text may contain archaic spelling, free word
    order, and compound words.
2. **Identify only phrases describing work, the performance of work, or work-related
     actions.** This includes performing, contributing to, or participating in work,
     occupation, service, or duty for a livelihood, regardless of whether
    compensation is provided.
3. **Normalise each identified phrase into modern base form (infinitive)**
   - Use modern spelling and grammar.
   - Formulate each phrase as an infinitive verb phrase, even if the original phrase
       is not verbal..
4. **Return a list of normalised phrases** in the order they appear in the text.

# Output Format:
Return only a JSON object with the key **"normalised_phrases"** and a list of
    phrases in normalised base form. Return the JSON exactly in this format, without
     any additional text, comments, summaries, extra keys, or code block markers:

{
  "normalised_phrases": ["grind grain", "to be in service"]
}

# Examples
(Only short excerpts are shown in the examples. In the actual task, the entire text
    should be processed.)

### Input 1:
"/.../ seeing as hops have been so scarce of late? She answered that she had
    purchased hops from Tidö, to which she had made a small addition of some
    Wormwood Twigs, /.../"

### Output 1:
{
  "normalised_phrases": ["purchase hops"]
}

/.../

# Identify and normalise (in base form) the phrases describing work in the following
     petition:
```

Prompt Task 4: Combined Extraction and Normalisation

Swedish Original

```
Du är expert på lingvistisk analys av svenska historiska texter. Din uppgift är att
    analysera svenska suppliker från 1700-talet och identifiera alla fraser som
    beskriver arbete, arbetsutförande eller arbetsrelaterade handlingar. Först ska
    du extrahera dessa fraser exakt som de förekommer i texten, därefter ska du
    normalisera varje extraherad fras till modern grundform (infinitiv).

# Huvudinstruktioner
1. **Läs texten noggrant.** Texten kan innehålla gammaldags stavning, fri ordföljd
    och sammanskrivningar.
2. **Extrahera endast fraser som beskriver arbete, arbetsutförande eller
    arbetsrelaterade handlingar.** Detta inkluderar att utföra, bidra till eller
    delta i arbete, syssla, tjänst eller plikt för försörjning, oavsett om ersä
    ttning ges.
3. **Bevara originaltexten exakt vid extraktion.** Lägg INTE till, ta bort eller ä
    ndra ord, stavning eller skiljetecken i de extraherade fraserna.
4. **Normalisera varje extraherad fras till modern grundform.**
   - Använd modern stavning och grammatik.
   - Formulera varje fras som en verbfras i infinitiv, även om originalfrasen inte ä
       r verbal.
5. **Bevara ordningen.** Den normaliserade frasen ska motsvara den extraherade
    frasen på samma position i listan.

# Outputformat
Returnera endast ett JSON-objekt med nycklarna **"extracted_phrases"** (fraser exakt
     ur texten) och **"normalised_phrases"** (motsvarande fraser i normaliserad
    grundform). Listorna ska ha samma längd och ordning. Returnera JSON exakt i
    detta format, utan extra text, kommentarer, sammanfattningar, extra nycklar
    eller kodblock-markeringar:

{
  "extracted_phrases": ["mala någon säd", "i tiänst warit"],
  "normalised_phrases": ["mala säd", "vara i tjänst"]
}

# Exempel
(Endast korta utdrag visas i exemplen. I uppgiften ska hela texten bearbetas.)

### Input 1:
"/.../ efter som thenne tiden warit så swårt efter humble? Swarades hon hade kiöpt
    humbla ifrån Tidön, hwartil hon giordt en liten tilökning af några Malörts
    Qwistar, /.../"

### Output 1:
{
  "extracted_phrases": ["kiöpt humbla"],
  "normalised_phrases": ["köpa humle"]
}

# Identifiera, extrahera och normalisera de fraser som beskriver arbete i följande
    supplik:
```

English Translation

```
You are an expert in the linguistic analysis of Swedish historical texts. Your task
    is to analyse Swedish petitions from the 18th century and identify all phrases
    describing work, the performance of work, or work-related actions. First, you
    must extract these phrases exactly as they appear in the text; subsequently, you
     must normalise each extracted phrase into modern base form (infinitive).

# Main Instructions
1. **Read the text carefully.** The text may contain archaic spelling, free word
    order, and compound words.
2. **Extract only phrases describing work, the performance of work, or work-related
    actions.** This includes performing, contributing to, or participating in work,
    occupation, service, or duty for a livelihood, regardless of whether
    compensation is provided.
3. **Preserve the original text exactly during extraction.** Do NOT add, remove, or
    change words, spellings, or punctuation marks in the extracted phrases.
4. **Normalise each extracted phrase into modern base form**
   - Use modern spelling and grammar.
   - Formulate each phrase as an infinitive verb phrase, even if the original phrase
       is not verbal.
5. **Preserve the order.** The normalised phrase must correspond to the extracted
    phrase at the same position in the list.

# Output Format:
Return only a JSON object with the keys **"extracted_phrases"** (phrases exactly
    from the text) and **"normalised_phrases"** (corresponding phrases in normalised
     base form) and a list of phrases in normalised base form. The lists must have
    the same length and order. Return the JSON exactly in this format, without any
    additional text, comments, summaries, extra keys, or code block markers:

{
   "extracted_phrases": ["grind some grain", "in service have been"],
   "normalised_phrases": ["grind grain", "to be in service"]
}

# Examples
(Only short excerpts are shown in the examples. In the actual task, the entire text
    should be processed.)

### Input 1:
"/.../ seeing as hops have been so scarce of late? She answered that she had
    purchased hops from Tidö, to which she had made a small addition of some
    Wormwood Twigs, /.../"

### Output 1:
{
  "extracted_phrases": ["purchased hops"],
  "normalised_phrases": ["purchase hops"]
}

/.../

# Identify, extract and normalise the phrases describing work in the following
    petition:
```