Leveraging Cross-Verification to Enhance Zero-Shot Prompting for Care Document Data Extraction

Laura Steffny[a], Nanna Dahlem, Robert Becker[b] and Dirk Werth[c]
August-Wilhelm Scheer Institute for Digital Products and Processes gGmbH, Saarbrücken, Germany
{*laura.steffny, nanna.dahlem, robert.becker, dirk.werth*}*@aws-institut.de*

Keywords: Zero-Shot Prompting, Cross-Verification, Chain-of-Verification, CoV, Large Language Model, LLM, Data Extraction, Care, Documentation.

Abstract: Automating care documentation through artificial intelligence (AI), particularly using large language models (LLMs), has great potential to improve workflow and efficiency in healthcare applications. However, in clinical or care environments where errors can have serious consequences, ensuring the reliability and accuracy of LLM output is essential. Zero-shot prompting, an advanced technique that does not require task-specific training data, shows promising results for data extraction in domains where large, well-structured datasets are scarce. This paper investigates how cross-verification affects zero-shot prompting performance in extracting relevant care indicators from unbalanced nursing documentation. The extraction was evaluated for three indicators on a dataset of care documentation from 38 participants across two facilities. The results show cross-verification significantly improves extraction accuracy, particularly by reducing false positives. While term extraction alone achieved around 80% accuracy, at lower temperature settings (0.1) cross-verification increased accuracy to 96.74%. However, cross-verification also increased missed terms when no corresponding sentences were found, even though terms were in the ground truth. This study highlights the potential of cross-verification in care documentation and offers suggestions for further optimization, especially with unstructured text and unbalanced data.

1 INTRODUCTION

Artificial Intelligence (AI) holds the potential to significantly improve healthcare by automating administrative tasks, supporting diagnostics, and optimizing patient care (Beck et al., 2023). In the context of nursing, AI applications could alleviate the burden on healthcare professionals, addressing staffing shortages while enhancing job satisfaction. AI shows significant potential in improving clinical documentation, a task that remains both time-consuming and essential to healthcare operations. The automation of these processes through AI, particularly using Large Language Models (LLMs), can streamline workflows and improve efficiency (Zernikow et al., 2023).

LLMs have demonstrated their ability to assist in various healthcare tasks, ranging from automating administrative duties to generating patient information and supporting clinical decision-making (Treder et al., 2024) (Zernikow et al., 2023). However, despite their potential, challenges remain in integrating LLMs into healthcare settings, particularly regarding data privacy, security, and ethical implications (Park et al., 2024). Moreover, the issue of ensuring that LLMs operate responsibly and with human oversight is critical, since errors in clinical environments can have severe consequences (Sonntagbauer et al., 2023).

One emerging solution to improve the accuracy and reliability of LLMs is prompt engineering, a technique that involves carefully designing inputs to guide the model toward generating desired outputs. One of the techniques in this field is zero-shot learning, where LLMs are used to extract relevant information without needing prior task-specific example data (Russe et al., 2024). This method is particularly relevant in nursing documentation, where large volumes of data must be processed efficiently while maintaining accuracy (Sellemann, 2021). However, ensuring the precision of zero-shot prompting remains a significant challenge, necessitating new approaches to enhance its performance in real-world applications.

This research explores how cross-verification techniques can enhance the effectiveness of zero-shot

[a] https://orcid.org/0009-0003-0014-8590
[b] https://orcid.org/0000-0002-7692-941X
[c] https://orcid.org/0000-0003-2115-6955

Steffny, L., Dahlem, N., Becker, R. and Werth, D.
Leveraging Cross-Verification to Enhance Zero-Shot Prompting for Care Document Data Extraction.
DOI: 10.5220/0013177600003911

In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Volume 2: HEALTHINF*, pages 115-128
ISBN: 978-989-758-731-3; ISSN: 2184-4305

prompting for data extraction in nursing documentation.

Research Question: *How does cross-verification mitigate the challenges of zero-shot prompting in extracting relevant nursing terms from unbalanced nursing documentation datasets?*

The paper is structured as follows: Chapter 2 discusses the relevant literature in the area of prompt engineering methods and data extraction using LLMs. Both classical and modern approaches are reviewed, particularly in the context of zero-shot prompting and the use of cross-verification. Chapter 3 describes the data set and the requirements arising from care practice and the regulatory framework. Chapter 4 presents the study design, including the methodology for term extraction and the implementation of cross-verification. The results of the study are analysed and discussed in Chapter 5, with a particular focus on the different types of error and the impact of the different verification methods. Finally, Chapter 6 highlights the limitations of the study and possible approaches for future work.

2 RELATED WORK

2.1 Advancements in Term Extraction

Term extraction, a subset of information extraction, is a critical task in natural language processing (NLP) that focuses on identifying and classifying key terms from text. This process is essential for tasks such as information retrieval, machine translation, and knowledge discovery (Mansouri et al., 2008).

Early techniques, including Maximum Entropy models (Chieu and Ng, 2003) and Hidden Markov Models (Zhou and Su, 2001), were effective in identifying named entities across various domains but require extensive manual feature engineering and struggle with adapting to new domains. With the advent of deep learning, techniques such as Bidirectional LSTM-CNNs (Chiu and Nichols, 2016) have achieved state-of-the-art results on named entity recognition (NER) tasks. However, challenges such as domain portability, handling nested entities, and ensuring consistent performance across languages remain (Yu et al., 2020).

More recently, LLMs are used for term extraction, particularly in specialized domains such as biomedicine. LLM-based systems, like those developed for biomedical literature (Monajatipoor et al., 2024), demonstrate the effectiveness of prompt engineering in improving performance, particularly in low-resource scenarios. Approaches such as NuNER and GPT-NER (Wang et al., 2023b) (Bogdanov et al., 2024) have transformed traditional sequence labeling tasks into text generation tasks, showing significant promise, especially when external knowledge is integrated (Bian et al., 2023). LLMs still face challenges such as hallucination, where models generate incorrect or irrelevant information, and gaps in domain-specific knowledge (Wang et al., 2023b). To address these issues, researchers are exploring strategies like adversarial training and external resource integration to enhance model robustness (Jin et al., 2023) (Monajatipoor et al., 2024). Reconfiguration of NER tasks from sequence labeling to text generation further improves performance in complex domains (Wang et al., 2023b). Hybrid approaches, that combine traditional statistical methods with machine learning models, seek to balance the precision of rule-based methods with the adaptability and scalability of machine learning (Yuan et al., 2017). LLMs like GPT-3.5 have demonstrated high performance in domain-specific term extraction tasks (Chataut et al., 2024) (Giguere, 2023).

LLMs have shown particular promise in low-resource environments (Deng et al., 2022) but still face challenges such as model bias, robustness, resource requirements, and concerns around transparency, privacy and responsible AI (Li et al., 2024). Nevertheless, LLMs offer a promising avenue for improving the precision and scalability of term extraction across various domains, including healthcare, legal frameworks, and education (Ding et al., 2023).

2.2 LLMs and Imbalanced Datasets

Imbalanced datasets pose a significant challenge in NLP, especially in tasks such as text classification, NER, and information extraction. These datasets are characterized by a disproportionate distribution of class labels, where some classes are underrepresented. As a result, machine learning models, including LLMs, often struggle to correctly predict the minority classes, leading to biased and less accurate results (Cloutier and Japkowicz, 2023). In recent years, extensive research has focused on addressing these challenges, particularly in the context of LLMs, which are known for their ability to handle large-scale text data.

Developed approaches to mitigate effects of imbalanced data in LLMs include transfer learning and fine-tuning. Associated methods such as modified weighting strategies, particularly in multilingual models (Jung and van der Plas, 2024), and Deep One-Class Fine-Tuning (DOCFT) (Bose et al., 2023) have shown promise in enhancing performance. An-

other promising approach is LLM-based data augmentation, which involves generating synthetic data to provide a more diverse and balanced training set. LLMs, such as GPT-3, have been used to create synthetic samples that enrich underrepresented classes. This approach has been effectively applied in domains such as clinical NLP tasks, where unbalanced data is a common challenge (Cai et al., 2023). Prompt engineering has also emerged as a crucial strategy for improving LLM performance on imbalanced datasets. Studies have shown that well-designed prompts can significantly enhance the model's ability to generate accurate predictions for imbalanced data (Kochanek et al., 2024). Looking forward, the integration of already mentioned few-shot learning techniques with LLMs combined with data augmentation and prompt engineering, has shown great potential in improving model resilience against imbalanced datasets (Billion Polak et al., 2024).

2.3 Prompt Engineering

Prompt engineering has emerged as a crucial technique for optimizing the performance of LLMs across various domains, including healthcare (Meskó, 2023). This technique involves the careful design of inputs—known as "prompts"— that guide model outputs toward desired results. In healthcare, prompt engineering is increasingly being applied to support tasks, such as clinical documentation, by efficiently extracting and processing relevant information (Jiaqi et al., 2023).

Various approaches have been developed to improve the effectiveness of prompts. These include instruction-based, information-based, and reformulation prompts, each employing distinct strategies to provide models with clear and context-aware instructions (Rathod, 2024). The goal of these techniques is to enhance the accuracy and relevance of model responses. To further improve the effectiveness of prompt engineering, researchers have introduced systematic frameworks such as PE2 and CLEAR, which help optimize prompt clarity, conciseness, and context awareness (White et al., 2023) (Lo, 2023).

A key area where prompt engineering has gained attention is in the application of few-shot and zero-shot learning. Few-shot prompting enables models to perform tasks after being exposed to a limited number of examples, while zero-shot prompting requires no task-specific training data, making it highly suitable for scenarios where labeled data is scarce or unavailable (Reynolds and McDonell, 2021) (Zhou et al., 2022). Zero-shot learning has demonstrated significant potential in fields such as nursing documentation, where large datasets often contain unbalanced distributions of terms, and manual data labeling is impractical (Wang et al., 2023a). Despite the promise of zero-shot prompting, crafting effective prompts can be time-consuming and complex, requiring careful attention to detail to ensure that the model produces accurate and meaningful outputs (Wang et al., 2023a). Researchers have developed innovative techniques, such as inverse prompting and self-adaptive prompts, which improve the model's ability to generalize across tasks and handle complex multimodal data (Li et al., 2023) (Wang et al., 2023a). These methods have been particularly effective in refining zero-shot models for clinical environments.

Finally, ethical considerations such as bias and transparency are important aspects of prompt engineering. Addressing these concerns will require ongoing efforts and the establishment of ethical guidelines for AI in critical healthcare settings (Ahmed et al., 2024). As the field of prompt engineering continues to evolve, effective prompting will become an essential skill for leveraging LMMs full potential in healthcare and beyond (Lo, 2023).

2.4 Cross-Verification Techniques for LLMs

The increasing complexity and widespread use of LLMs have underscored the need for robust verification techniques to ensure the accuracy, reliability, and trustworthiness of these models. Recent research has explored several cross-verification methods to enhance LLM performance by addressing issues such as hallucination, reasoning accuracy, and factual consistency (Kang et al., 2023) (Dhuliawala et al., 2023). Cross-verification, in this context, involves using multiple independent processes or models to verify the outputs of an LLM, ensuring higher fidelity and reducing the likelihood of incorrect or misleading information. Prominent approaches like real-time verification and rectification, where verification steps are incorporated during the generation of text, reduce hallucinations by continuously validating outputs against established facts or external knowledge sources (Kang et al., 2023). Semantic-aware cross-checking techniques are used to detect hallucinations and inconsistencies in LLM outputs by comparing semantic information across different sections of generated content and the input prompt (Zhang et al., 2023). The MILL framework, which applies cross-verification in zero-shot query expansion by allowing LLMs to mutually verify their generated queries (Jia et al., 2023), ensures that the LLM-generated expansions are consistent and factual across multiple itera-

tions. Another key technique is chain-of-verification (CoVe), which ensures that an LLM's initial response undergoes subsequent verification stages to validate its correctness and consistency (Dhuliawala et al., 2023). CoVe has been particularly effective in improving factuality in complex tasks, such as question answering and reasoning, by employing a multi-step process that cross-checks model-generated outputs.

3 MATERIAL AND METHODS

3.1 Nursing Professional Requirements

As part of a preliminary study, 18 nursing professionals (age: 41.19 ± 11.30 years, work experience: 15.69 ± 9.58 years) from two long-term care facilities were surveyed regarding their views on what an optimal intervention component for care planning should look like. The methodology used was *Cultural Probes*, where the nursing professionals could select various components from the areas of *intervention description*, *reasons for the intervention*, and *representation of recommendation level* (see Figure 1).

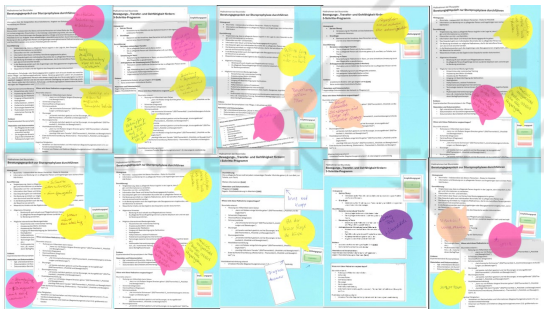

Figure 1: Excerpt from the results of the cultural probes preliminary study.

For the design of the presented system, the component *reasons for the intervention* was of particular importance. Participants were provided with four different presentation formats, ranging from a very compact information representation to a detailed elaboration. Overall, 69.45% of the participants preferred the most detailed version. Additionally, participants could leave comments on their selection using Post-it notes (see Figure 1). Thematic analysis of participant responses identified three key areas: the importance of clear and structured presentation, the necessity of integrating nurse documentation excerpts, and a preference for concise yet detailed descriptions of care measures and topics. Based on the findings, two functional requirements for the approach were established:

- Extraction of defined values for specific indicators, even if these values do not exactly match the notation used in the source text.
- Extraction of verbatim quotations from the text source corresponding to each specific value identified.

3.2 Language Model Requirements

Besides the nursing professional requirements, the extraction approach must satisfy three core criteria: language applicability, data privacy, and computational feasibility.

In terms of language applicability, the following points were of particular importance: The model was required to accommodate German texts, including those pertaining to nursing care terminology, given that the dataset comprised documents from German care facilities. The second requirement was data privacy. In light of the necessity for GDPR compliance, cloud-based models such as Chat-GPT4 were deemed unsuitable. It was imperative that an on-premises model be utilized to guarantee comprehensive control and data security. The model's size and computational requirements were also taken into consideration. The model had to strike a balance between performance and computational efficiency. It needed to be sufficiently large to accommodate complex language processing, yet still feasible for local deployment within hardware limits.

3.3 Dataset Description

The dataset used in this study was collected as part of the ViKI pro research project (grant number 16SV8870) and consists of nursing documentation data from 38 participants across two nursing facilities. The data were manually pseudo-anonymized by the quality managers of these facilities, with identifiable information such as the first and last names of residents and their relatives removed. Each resident gave their written consent.

The nursing documentation data used for evaluation were derived from the "Structured Information Collection" (Strukturierte Informationssammlung, SIS). The SIS integrates the self-assessment of the care recipient with the professional assessment of the caregiver. The SIS covers the following domains: What is on your mind? What can we do for you? What brings you to us? (Topic 0), Cognitive and communicative abilities (Topic 1), Mobility and physical agility (Topic 2), Disease-related requirements and burdens (Topic 3), Self-care (Topic 4), and Social relationships and interactions (Topic 5). For inpatient

care, an additional category, "Living/Domestic Environment (Topic 6)" is included. Furthermore, the care recipient is asked initially about their current concerns or needs, resulting in a total of seven domains (see example in Figure 3).

A total of 266 SIS topics were evaluated. The average number of words in the topic areas varied from 12.55 ± 17.85 for topic area 6 to 105.66 ± 46.55 for topic area 4 (see Table 1).

Table 1: Average number of words in the SIS, broken down by subject area.

SIS topic	Word Count {Mean ± Std}
0	25.21 ± 23.79
1	88.89 ± 54.29
2	98.24 ± 53.26
3	69.94 ± 39.21
4	105.66 ± 46.55
5	29.11 ± 32.24
6	12.55 ± 17.85

4 STUDY DESIGN

Based on the requirements derived in Chapter 3.1, the extraction of predefined values and corresponding literal citations from the subject areas of the SIS for specific indicators was established as a core objective of this study. Figure 2 illustrates the desired outcomes derived from the needs of caregivers.

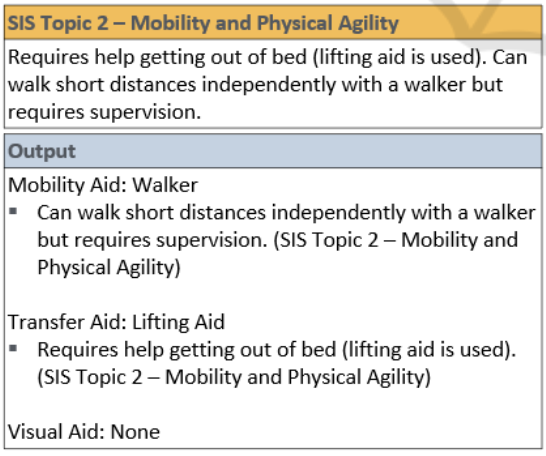

Figure 2: Presentation of the desired outcomes derived from the needs of the caregivers.

To achieve the desired outcome, a system was designed leveraging LLMs in combination with traditional post-processing techniques in Python. The proposed approach comprises five key components, which together ensure both the extraction and validation of data to meet the specified requirements (see Figure 3). The components of the system are as follows:

1. A language model that extracts terms from the SIS subject areas and maps them to predefined notations.
2. A Python-based post-processing pipeline that validates and corrects the language model's output to ensure data format consistency.
3. A language model that retrieves the literal citations from the SIS subject areas corresponding to the extracted terms.
4. A cross-verification pipeline implemented in Python to ensure that each extracted term has an appropriate corresponding literal citation.
5. A Python function that formats the final verified results and delivers them to the caregiver in the required format.

This methodology was applied to analyze three key indicators: mobility aid, transfer aid, and visual aid. Each indicator has specific term keys associated with possible values, as outlined in Table 2.

Table 2: Overview of the analyzed indicators with their term keys and values (overview in English, used in the original prompt in German).

Indicator Name	Term Key	Value Options
Mobility Aid	walk	Walking Stick, Crutches, Rollator Wheelchair, E-Wheelchair None
Transfer Aid	transfer	Transfer Belt, Turntable, Slide Board, Slide Mat, None
Visual Aid	see	Eyeglasses, Reading Glasses, Magnifier, Contact Lenses, Other, None

4.1 Formulation of Prompts

As described in the previous chapter, the system consists of two language models. For each language model, a system prompt and a user prompt are defined. The system prompt provides fixed instructions created by the developers to constrain the responses of the LLM, specifically in terms of scope, task, context, and style. The user prompt, in contrast, is the request made by the user to the LLM.

4.1.1 Term Extraction

The **Term Extraction** task utilized zero-shot prompt engineering, where a carefully designed system prompt and user prompt were employed to extract terms related to specific indicators. The system

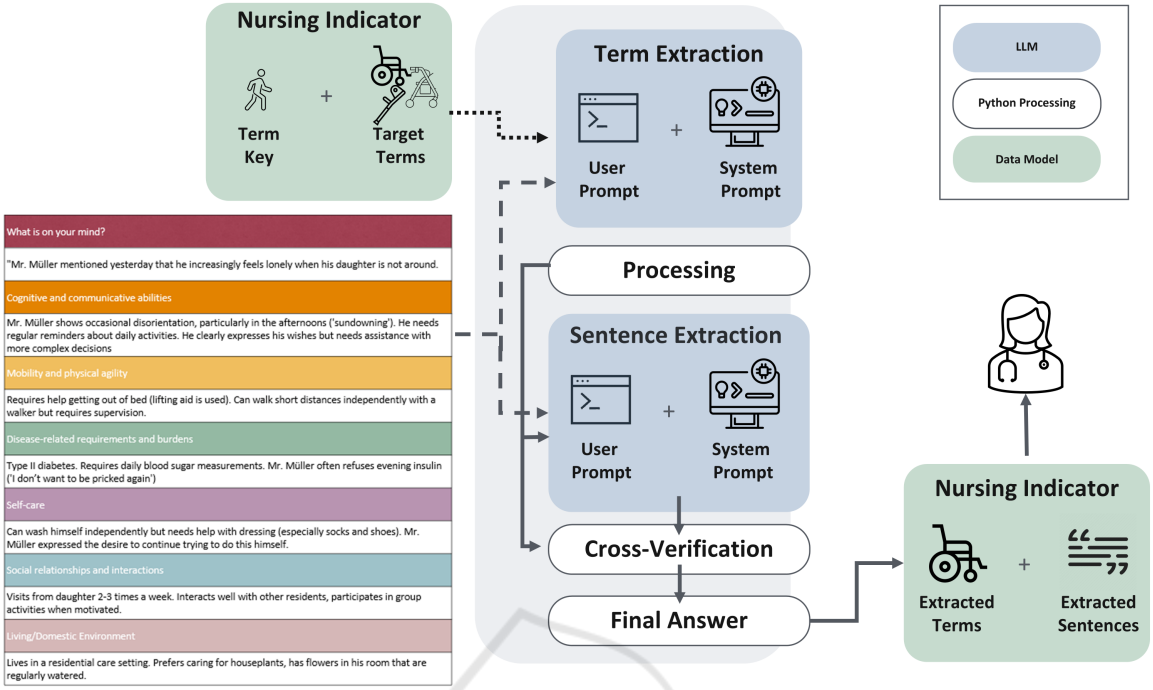

Figure 3: Schematic representation of the presented approach.

prompt instructed the model to act as an assistant, answering questions strictly based on explicit information available in the text.

The system prompt was developed iteratively through a process of trial and adjustment. Various prompts were tested based on the *plan-and-solve (PS) prompting* approach, as proposed by Wang et al. (2023). PS prompting is a zero-shot method designed to enhance the reasoning capabilities of LLMs, particularly for solving complex, multi-step tasks. The method instructs the LLM to first devise a plan, breaking the task into smaller subtasks, followed by executing each subtask according to the plan. After each test run, difficulties encountered with the prompt were evaluated and necessary adjustments were made.

Ultimately, a system prompt consisting of five key components was finalized (see Figure 4), as described below:

1. **Role Specification:** This section defines the role of the LLM. By assigning the role of an ”assistant,” the model is directed to provide structured and context-aware responses. This ensures the model responds to prompts while focusing on specific contexts and maintaining consistency.
2. **Response Format:** The model is instructed to provide answers in a multiple-choice format, with the possibility of selecting more than one correct answer. This format aligns with the study's methodology, where certain indicators (e.g., mobility aids) may have multiple correct answers based on the context. Additionally, the response is required to be formatted in JSON with two keys: `correct_answer_id` (a list of integers representing selected answer IDs) and `correct_answer_str` (a list of corresponding answer strings). This structured format facilitates machine-readability and efficient post-processing.
3. **Strictness:** The model is explicitly instructed to base its responses only on information that is clearly mentioned in the text. Assumptions or interpretations beyond the provided content are not allowed. This restriction ensures that the model adheres to factual extraction, avoiding any speculative or creative responses.
4. **Handling Absence of Relevant Information:** When no relevant information (such as mobility aids or other indicators) is found in the context, the model is required to select 'None'. This ensures that missing data is explicitly accounted for, preventing the model from making incorrect assumptions when information is absent.
5. **Justification with Explanation:** After producing the JSON output, the model must include a justification paragraph starting with the word ”Explanation.” This paragraph should clearly explain the reasoning behind the selected answers, providing transparency and insight into the decision-making process based on the context provided in the text.

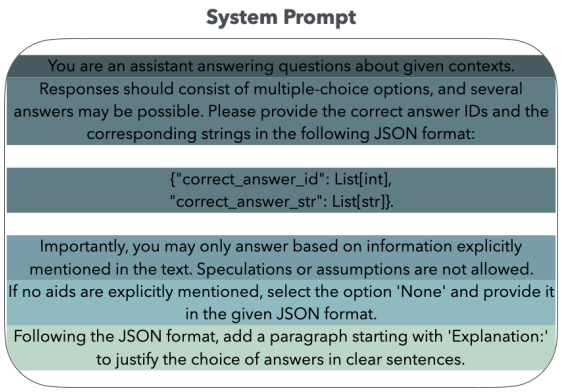

Figure 4: Final system prompt that was used for term extraction. Note: Prompt was translated into English, in the original language being German.

In addition to the system prompt, a user prompt is used to directly query the language model for specific information. The user prompt can also be broken down into five components.

1. **Context Introduction:** This part of the prompt introduces the specific context for the language model by referencing the `sis_topic`, which provides a description of the resident for one SIS subject.
2. **Targeted Question:** The model is instructed to focus on a specific category of aids, as indicated by the `term_key`. This key is dynamically substituted with terms such as ”walk,” ”see,” or ”transfer” (see Table 2), depending on the specific indicator being queried.
3. **Answer Options:** The model is provided with a predefined set of possible answer options related to the indicator. For example, for mobility aids, the options might include ”walking stick,” ”crutches,” or ”none” (see Table 2). By giving the model these specific choices, the output is constrained to these values, making the response more structured and easier to process.
4. **Step-by-Step Reasoning:** This instruction prompts the LLM to engage in step-by-step reasoning. Encouraging the model to take this structured approach to problem-solving improves its performance on more complex, multi-step tasks by forcing it to break down the task and consider the evidence more carefully.
5. **Focusing on Explicit Information:** This final instruction reinforces the constraint that the language model should only extract terms that are explicitly mentioned in the resident’s description. It ensures that the model doesn’t speculate or infer beyond what is clearly stated in the text.

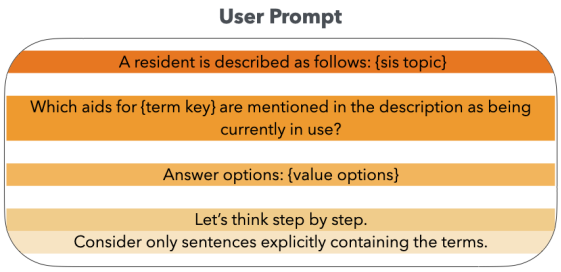

Figure 5: Final user prompt that was used for term extraction. The placeholder are described in Table 2. Note: Prompt was translated into English, in the original German.

4.1.2 Processing

The **Processing** function consisted of a post-processing pipeline that verifies the output of the term extraction LLM. If the output was either not in valid JSON format or returned an empty list, the pipeline automatically filled the list with the value `["None"]`. This approach ensured that incomplete responses from the model were systematically addressed, reducing the likelihood of false negatives.

4.1.3 Sentence Extraction

The **Sentence Extraction** system prompt employs a comparable structure to that of the term extraction system, comprising the following components: (1) Role Specification, (2) Response Format, (3) Strictness, and (5) Justification with Explanation. In contrast to the extraction term, no instructions are provided regarding the handling of missing information. In this case, an empty list is to be returned, rather than a specific value (see Figure 6).

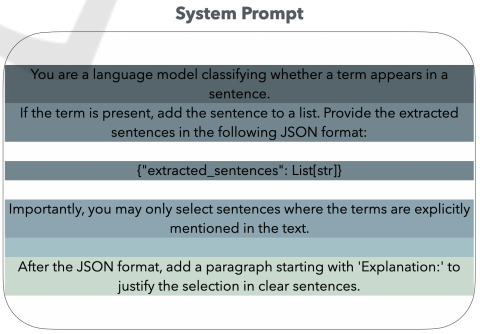

Figure 6: Final system prompt that was used for sentence extraction. Note: Prompt was translated into English, in the original German.

The user prompt employs the same scheme as the one used for term extraction, with slight adaptations to the wording. The only difference is that the answer options have been replaced with the output values of the processing function (see Figure 7).

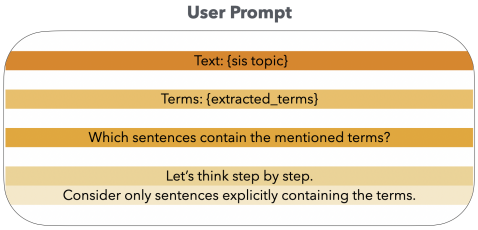

Figure 7: Final user prompt that was used for sentence extraction. Note: Prompt was translated into English, in the original German.

4.1.4 Cross-Verification

The **Cross-Verification** function enhances the precision of the results through a sentence-level verification process that leverages the output of the sentence extraction function. In the absence of explicit sentences containing the extracted term, the corresponding value is reset to `"None"`. This verification step introduces an additional layer of accuracy to the extraction process.

4.2 Running of Prompts

To meet the outlined requirements, the multilingual state-of-the-art Llama 3.1-8B model was selected. It was chosen specifically for its strong performance in handling diverse linguistic challenges, including the German language. Despite the relatively smaller proportion of German data, the model's extensive training ensures a sufficient understanding of the nuances required for interpreting German, including domain-specific terminology used in nursing care. Additionally, the model's architecture allowed for flexible deployment, including local execution, which was essential for meeting GDPR compliance. To address the computational requirements, the model was configured to operate in a lower precision 4-bit mode, reducing the GPU memory requirements from 38.4 GB to approximately 4.8 GB. This configuration enabled efficient operation while retaining the performance necessary for the accurate extraction of terms from nursing documentation.

The formulated prompts were executed in a self-written Python script utilizing the transformer pipeline of Hugging Face. The model id `meta-llama/Llama-3.1-8B-Instruct`, with its default parameters, was used. The sole exception was the `temperature` parameter, for which a series of values were evaluated during the course of the experiment.The choice of temperature parameters plays a crucial role in adjusting the behavior of language models, particularly in controlling the diversity and creativity of generated outputs. In this study, temperatures were varied to determine the optimal value for extracting nursing indicators. For term extraction, temperatures of 0.1, 0.3, and 0.5 were tested. Lower temperatures (especially 0.1) were chosen to minimize output variability and maximize precision, as the model focuses more on the given context and is less likely to generate creative or "hallucinated" terms. For sentence extraction, a broader range of temperatures (0.1, 0.3, 0.5, 0.7, and 1.0) was tested to explore how model creativity impacts sentence verification. Higher temperatures (e.g., 0.7 or 1.0) were included to help the model identify alternative or implicit formulations that may not exactly match the extracted terms but carry the same meaning, particularly in cases where subtle or indirect references are present in the text.

4.3 Analysis of Prompts

Chapter 5 analyses the results of the term extraction experiments under different temperature settings (see Chapter 4.2) with and without cross-verification. First, the effect of temperature on the accuracy of term extraction is analysed. Different temperature parameters are compared in order to identify the optimal conditions for the most accurate extraction. The error categories described in Table 4 are then analysed in detail. This analysis includes the most common types of errors that occurred during term extraction and their distribution under the different temperature settings. The influence of the processing step (see also Figure 3) on the reduction of these errors is then analysed. The extent to which static post-processing has improved the consistency and accuracy of the results, irrespective of the temperature settings used, is assessed. Finally, the effect of cross-verification is analysed. Here, the temperature for term extraction was set to the determined optimum (0.1), while different temperatures were used for sentence extraction. The effect of cross-verification, which is strongly dependent on the output of the sentence extraction (see Figure 3), is evaluated in terms of the improvement in overall accuracy and the reduction in false-positive extractions.

5 FINDINGS AND DISCUSSION

In this study, the results were evaluated in comparison to a manually annotated benchmark dataset. For each of the three processing steps that directly affected the final response of the extracted terms (term extraction, processing and cross-verification), the number of correctly extracted terms was calculated as a proportion

Table 3: Resulting error categories. The table presents different error cases that can occur when an LLM extracts terms from texts and compares them to the Ground Truth.

Error Category (Abbreviation)	Description	Output Method	Ground Truth
'None' extracted but term in Ground Truth (NETGT)	A term was not extracted, although it is present in the Ground Truth.	["None"]	["Wheelchair"]
Term extracted but Ground Truth is 'None' (TEGTN)	A term was extracted, although it is not present in the Ground Truth.	["Wheelchair"]	["None"]
Missing Values (MV)	One or more terms from the Ground Truth are missing in the extraction.	["Wheelchair"]	["Wheelchair", "Walker"]
Too Many Values (TMV)	More terms were extracted than are present in the Ground Truth.	["Wheelchair", "Walker"]	["Wheelchair"]
Empty List (EL)	An empty list was returned, although the Ground Truth contains the value "None".	[]	["None"]

of the total number of 266 sis topics. The benchmark dataset comprises an imbalanced distribution of the analyzed indicators and their corresponding values, as shown in Table 4.

Table 4: Overview of the number of occurring indicator terms per SIS subject area.

	SIS Topics						
	0	**1**	**2**	**3**	**4**	**5**	**6**
None	112	90	78	113	113	113	114
Rollator	2	1	24	0	0	0	0
Wheelchair	0	1	17	0	1	1	0
E-Wheelchair	0	0	1	0	0	0	0
Crutches	0	0	1	0	0	0	0
Walking Stick	0	0	3	0	0	0	0
Eyeglasses	0	17	0	1	0	0	0
Reading Glasses	0	5	0	0	0	0	0
Magnifier	0	1	0	0	0	0	0
Slideboard	0	0	2	0	0	0	0

5.1 Term Extraction and Processing

5.1.1 Accuracy Analysis at Different Temperature Values

Figure 8 (top) shows the accuracy analysis results for aid extraction at different temperature values. The analysis focuses on Transfer Aids, Mobility Aids, and Visual Aids indicators, with the highest accuracy achieved at a temperature of 0.1.The highest accuracy values were achieved at a temperature of 0.1 across all indicators, particularly for Visual Aids. For Transfer and Mobility Aids, accuracy decreased as the temperature increased, while Visual Aids maintained relatively high accuracy across all temperature settings. Figure 8 illustrates these trends in detail. These observations are corroborated by the overall accuracies across all indicators, as illustrated in Table 5. Processing consistently shows higher accuracy values compared to term extraction, especially at higher temperatures.

Figure 8 (bottom) illustrates the accuracy achieved in processing the identical indicators and temperatures, as previously demonstrated in the analysis of results pertaining to term extraction. Subsequently, the results are compared with those of the term extraction in order to identify any improvements. Mobility, and Visual Aids indicators shows that processing consistently outperforms term extraction across all temperature settings. For all indicators, accuracy is highest at 0.1, with improvements in processing accuracy ranging from slight increases at lower temperatures to more significant gains at higher temperatures. Visual Aids consistently demonstrated the highest accuracy overall. The average accuracies of processing and term extraction for all indicators, as summarized in Table 5.

Table 5: Overall accuracy of term extraction and processing at different temperature settings (0.1, 0.3 and 0.5).

	0.1	0.3	0.5
Term Extraction	80.95	75.56	71.55
Processing	81.83	80.45	78.57

5.1.2 Analysis of Error Types and Processing Effects

The error analysis of term extraction and processing was carried out using the indicators defined in Table 2 and the error categories listed in Table 3 were used to analyse the errors. The analysis includes the most common types of errors that occurred during the experiments and compares their occurrence at different

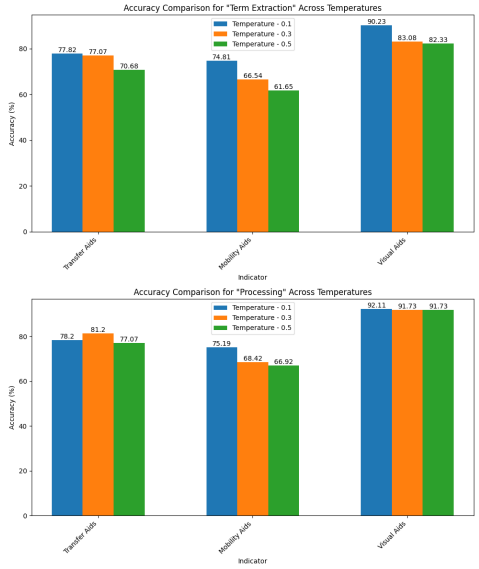

Figure 8: Comparison of the accuracy of term extraction and processing at three different temperatures (0.1, 0.3, 0.5) for the Transfer, Mobility and Visual Aids indicators.

temperature settings (0.1, 0.3 and 0.5) for term extraction and after processing. NETGT does not occur at the lower temperatures (0.1), but occurs sporadically at the higher temperatures (0.3 and 0.5). This error remains in the processing as the step does not explicitly validate the extraction. TEGTN is the most common error across all temperature settings. Here the model extracts terms that are not present in the ground truth. It is clear that the error increases at higher temperatures (0.5). The processing step has no effect on this error as the incorrect extractions are not corrected. MV remains constant over all temperatures, as the model correctly extracts most of the values in the analysed cases, but does not identify individual terms. In these cases too, processing does not lead to any improvement. TMV occurs only at higher temperatures (0.5) and indicates an increasing uncertainty of the model when it extracts too many terms. Processing cannot eliminate this error either. EL, where the model returns an empty list instead of 'None', increases significantly with increasing temperature. The error rarely occurs at a temperature of 0.1, but more frequently at 0.3 and especially at 0.5. In processing, this error is completely corrected by systematically filling in missing terms with 'None'.

Overall, the error analysis shows that higher temperatures have a negative impact on the accuracy of term extraction, as the number of misclassifications (especially TEGTN and EL) increases with increasing temperature. The processing specifically targets the elimination of the EL category and can completely correct it.

Table 6: Error analysis for "Term Extractor" and "Processes" compared to "Ground Truth" across different temperature settings (0.1, 0.3, 0.5).

Error Type	Method	0.1	0.3	0.5
NETGT	Term Extractor	-	1	1
	Processes	-	1	1
TEGTN	Term Extractor	133	143	157
	Processes	133	143	157
MV	Term Extractor	8	8	8
	Processes	8	8	8
TMV	Term Extractor	-	-	4
	Processes	-	-	4
EL	Term Extractor	7	39	56
	Processes	-	-	-

5.2 Cross Verification

5.2.1 Accuracy Analysis at Different Temperature Values

Figure 9 and Table 7 present the accuracy analysis of cross-verification at various temperatures for the Transfer Aids, Mobility Aids, and Visual Aids indicators. All three indicators maintained high accuracies across all temperatures. The overall accuracy improved significantly with cross-verification, reaching up to 96.74%, a significant improvement over the term extraction and processing results, which averaged at 80.20% and 81.58%, respectively.

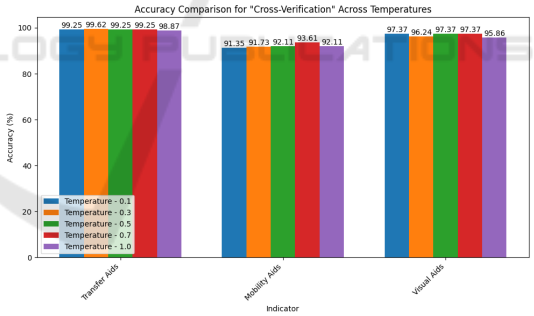

Figure 9: Comparison of cross-verification accuracies at varying sentence extraction temperatures (0.1, 0.3, 0.5, 0.7 and 1.0) and a term extraction temperature of 0.1 for the Transfer Aids, Mobility Aids and Visual Aids indicators.

Table 7: Overall accuracy of "Term Extractor" (TE), "Processes" (P), and "Cross-Verification" (CV) at different temperatures. The values for term extraction and processing marked with * were used with a fixed temperature of 0.1 for all cross-verification approaches.

	0.1	0.3	0.5	0.7	1.0
TE	80.95*	80.70*	81.45*	80.20*	81.58*
P	78.20*	81.70*	82.58*	81.58*	82.33*
CV	95.99	95.86	96.24	96.74	95.61

5.2.2 Analysis of Error Types and Cross-Verification Effects

The comparative analysis of errors in cross-verification, term extraction and processing is presented in Table 8. It is important to highlight that the temperature values shown relate to the Sentence Extractor and therefore directly influence cross-verification, as this method makes a correction based on the extracted sentences. It should be noted that NETGT does not occur in term extraction and processing. However, cross-verification shows that this error occurs at all analysed temperatures of the Sentence Extractor (0.1 to 1.0).

In contrast, TEGTN remains high across all temperatures in both term extraction and processing. Cross-verification reduces this error significantly. This suggests that cross-verification improves the ability to identify and correct erroneously extracted terms, especially at moderate temperature settings in the Sentence Extractor. MV is a relatively constant occurrence during the process of term extraction and processing, irrespective of temperature. In cross-verification, the Sentence Extractor results in a slight reduction in this error. The occurrence of EL, where no terms were extracted, has been eliminated during cross-verification as a result of the implemented processing correction.

In conclusion, cross-verification has been demonstrated to result in a notable reduction in TEGTN errors and to play a contributory role in the reduction of MV errors (see Figure 10). However, it should be noted that NETGT errors do occasionally occur, which is not the case with the other aforementioned methods. Furthermore, the results demonstrate that there is no notable discrepancy in performance across different temperatures when utilising the Sentence Extractor.

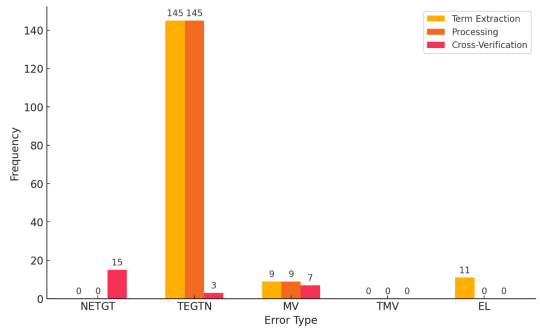

Figure 10: Number of errors per category for ‘Term Extractor’, ‘Processes’, and ‘Cross-Verification’ compared to ‘Ground Truth’ at a sentence extraction temperature of 0.7 and a fixed term extraction temperature of 0.1.

Table 8: Error analysis for ”Term Extractor” (TE), ”Processes” (P), and ”Cross-Verification” (CV) compared to ”Ground Truth” across varying sentence extraction temperatures (0.1, 0.3, 0.5, 0.7 and 1.0) and a term extraction temperature of 0.1.

Error Type	Method	0.1	0.3	0.5	0.7	1.0
NETGT	TE	-	-	-	-	-
	P	-	-	-	-	-
	CV	19	23	17	15	23
TEGTN	TE	133	133	139	145	140
	P	133	133	139	145	140
	CV	4	4	6	3	5
MV	TE	8	7	7	9	8
	P	8	7	7	9	8
	CV	7	5	5	7	6
EL	TE	7	8	7	11	6
	P	-	-	-	-	-
	CV	-	-	-	-	-

6 CONCLUSIONS

The objective of this study was to address the following research question: How does cross-verification mitigate the challenges of zero-shot prompting in extracting relevant nursing terms from unbalanced nursing documentation datasets? The findings demonstrate that cross-verification significantly enhances the accuracy of zero-shot prompting by reducing the number of false-positive extractions and increasing overall precision in the identification of key nursing indicators. In particular, cross-verification was shown to be an effective method for validating model outputs, particularly in the context of unstructured nursing documentation, where the lack of well-formed datasets presents a significant challenge for conventional extraction methods.

The analysis demonstrates that the incorporation of cross-verification enables the model to reduce the number of incorrect extractions, such as falsely identified terms, and to enhance the overall accuracy of the extraction process by up to 96.74%. However, the study also revealed that while cross-verification improved accuracy, it introduced a trade-off in that the likelihood of missed terms increased when no corresponding sentence was found, despite the terms being present in the ground truth. This indicates that the current form of cross-verification may be unduly restrictive in certain cases, particularly in the context of complex or implicit textual data.

In conclusion, this research presents a promising approach for automating care documentation using AI, particularly through the combination of zero-shot prompting and cross-verification. Although the method shows significant advances in data extraction from uneven nursing documentation, further en-

hancements are essential to address the constraints observed in handling implicit or absent sentence connections. Further research should concentrate on optimising cross-verification techniques and investigating methods of reducing the risk of missed extractions without compromising overall accuracy.

6.1 Limitations

While the combination of zero-shot prompting and cross-verification has yielded promising results, it is important to acknowledge the limitations of this approach.

The issue of data imbalance must also be addressed. The dataset was characterised by a significant prevalence of 'None' values, which resulted in the underrepresentation of certain key indicators. Although cross-verification proved effective in reducing false-positive extractions, it also increased the probability of failing to identify relevant terms when they were present in the ground truth. This suggests that the current implementation of cross-verification may be insufficient for handling rare or less frequent indicators in unbalanced datasets. One potential solution would be to augment the dataset with synthetic examples in order to provide better coverage of the terms that are underrepresented. The model exhibited constraints in its capacity to process text of varying degrees of complexity. It demonstrated a notable challenge in processing highly unstructured or complex nursing documentation, wherein terms were mentioned indirectly or not clearly linked to a specific sentence. This limitation was particularly evident in instances where sentence-level cross-verification was unable to discern implicit associations between terms and their corresponding textual passages. To address this limitation, the development of more sophisticated algorithms capable of understanding context beyond sentence boundaries may be required, such as semantic search or advanced contextual analysis. Although cross-verification proved beneficial for improving the precision of term extraction, its strictness at the sentence level resulted in an elevated number of omitted terms. In instances where an exact matching sentence could not be identified, despite the term being present in the text, the model rejected the extraction. This indicates that the existing approach to sentence-level cross-verification is unduly inflexible and may result in the loss of valuable information. Subsequent versions of this methodology may incorporate a more flexible verification process, potentially enabling the identification of approximate matches or the utilisation of multi-sentence context. The extent to which the findings can be generalised. The findings presented in this study are based on a dataset derived from German nursing documentation. It is therefore not possible to ascertain the extent to which the results can be generalised to other languages, domains or types of documentation. Further validation of the methodology in different contexts is required to confirm its broader applicability, particularly in settings with different documentation structures or regulatory requirements.

6.2 Future Work

In light of the findings and limitations of this study, a number of avenues for future research are proposed with a view to further refining and expanding the methodology presented.

The refinement of cross-verification techniques is a key objective. One of the principal challenges identified was the inflexibility of sentence-level cross-verification, which resulted in the omission of terms when no exact sentence match could be identified, despite their presence in the text. It would be beneficial for future research to concentrate on the creation of a more adaptable cross-verification strategy. This could entail context-based or semantic-level verification, whereby the model can verify terms by considering broader text passages or even multi-sentence context. The integration of semantic search algorithms or similarity-based matching techniques could enable the system to identify implicit relationships between terms and the corresponding sentences, thereby reducing the probability of missed terms. The potential for generalization to other domains is an avenue for future research. Although this study concentrated on German nursing documentation, it is vital to ascertain the viability of the proposed methodologies in other domains and languages. It would be beneficial for future research to extend the methodology to datasets from other fields, such as legal documents, technical manuals, or educational content, in order to assess the robustness and flexibility of the approach. Such an approach would not only validate the model's performance in diverse environments but also help to identify potential domain-specific challenges that may arise. Although zero-shot prompting demonstrated potential in this study, integrating few-shot learning techniques could further enhance the model's performance. The provision of a limited number of examples, particularly for ambiguous or less common terms, could assist the model in developing a more comprehensive understanding of the task and thereby enhance its extraction accuracy. The application of few-shot learning techniques could prove particularly advantageous in addressing the challenges posed by

complex text structures and rare terms, which were a significant limitation of the zero-shot approach. Further research could examine the potential of integrating few-shot examples into the extraction process in a dynamic manner, with the objective of achieving an optimal balance between the required training effort and the desired model improvement. Furthermore, the existing post-processing pipeline was primarily concerned with rectifying formatting inconsistencies and supplementing absent terms with the designation "None." Further research could examine more sophisticated post-processing techniques that extend beyond mere corrections. For instance, the incorporation of rule-based systems or secondary machine learning models could assist in the further refinement of extracted terms, through the verification of their consistency or cross-referencing with external knowledge sources. This may enhance the overall robustness and reliability of the term extraction process.

ACKNOWLEDGEMENTS

The study was part of the research projeject ViKI pro and funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF, grant number 16SV8870). We would like to thank the entire ViKI pro consortium (Deutsches Institut für angewandte Pflegeforschung e.V., Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Connext Communication GmbH, Johanniter Seniorenhäuser GmbH, Caritas-Betriebsführungs- und Trägergesellschaft mbH) for their collaboration, expertise, and support. We would also like to thank all study participants for their expertise and feedback on the design of the recommended measures.

REFERENCES

Ahmed, Mr., A., Hou, Prof. Dr., M., Xi, Dr., R., Zeng, Mr., X., and Shah, Dr., S. A. (2024). Prompt-Eng: Healthcare Prompt Engineering: Revolutionizing Healthcare Applications with Precision Prompts. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1329–1337. ACM.

Beck, S., Faber, M., and Gerndt, S. (2023). Rechtliche Aspekte des Einsatzes von KI und Robotik in Medizin und Pflege. *Ethik in der Medizin*, 35(2):247–263.

Bian, J., Zheng, J., Zhang, Y., and Zhu, S. (2023). Inspire the Large Language Model by External Knowledge on BioMedical Named Entity Recognition. *arXiv.org*.

Billion Polak, P., Prusa, J. D., and Khoshgoftaar, T. M. (2024). Low-shot learning and class imbalance: a survey. *Journal of Big Data*, 11(1).

Bogdanov, S., Constantin, A., Bernard, T., Crabbé, B., and Bernard, E. (2024). Nuner: Entity Recognition Encoder Pre-training via LLM-Annotated Data. *arXiv.org*.

Bose, S., Su, G., and Liu, L. (2023). *Deep One-Class Fine-Tuning for Imbalanced Short Text Classification in Transfer Learning*, pages 339–351. Springer Nature Switzerland.

Cai, X., Xiao, M., Ning, Z., and Zhou, Y. (2023). Resolving the Imbalance Issue in Hierarchical Disciplinary Topic Inference via LLM-based Data Augmentation. In *2023 IEEE International Conference on Data Mining (ICDM)*, volume 2011, pages 956–961. IEEE.

Chataut, S., Do, T., Gurung, B. D. S., Aryal, S., Khanal, A., Lushbough, C., and Gnimpieba, E. (2024). Comparative Study of Domain Driven Terms Extraction Using Large Language Models. *arXiv.org*.

Chieu, H. L. and Ng, H. T. (2003). Named entity recognition with a maximum entropy approach. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -*, volume 4, pages 160–163. Association for Computational Linguistics.

Chiu, J. P. and Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Cloutier, N. A. and Japkowicz, N. (2023). Fine-tuned generative LLM oversampling can improve performance over traditional techniques on multiclass imbalanced text classification. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE.

Deng, S., Ma, Y., Zhang, N., Cao, Y., and Hooi, B. (2022). Information extraction in low-resource scenarios: Survey and perspective.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., and Weston, J. (2023). Chain-of-Verification Reduces Hallucination in Large Language Models. *arXiv.org*.

Ding, Q., Ding, D., Wang, Y., Guan, C., and Ding, B. (2023). Unraveling the landscape of large language models: a systematic review and future perspectives. *Journal of Electronic Business & Digital Economics*, 3(1):3–19.

Giguere, J. (2023). Leveraging large language models to extract terminology. In Gutiérrez, R. L., Pareja, A., and Mitkov, R., editors, *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pages 57–60, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jia, P., Liu, Y., Zhao, X., Li, X., Hao, C., Wang, S., and Yin, D. (2023). Mill: Mutual Verification with Large Language Models for Zero-Shot Query Expansion. *North American Chapter of the Association for Computational Linguistics*.

Jiaqi, W., Enze, S., Sigang, Y., Zihao, W., Chong, M., Haixing, D., Qiushi, Y., Yanqing, K., Jinru, W., Huawen, H., Chenxi, Y., Haiyang, Z., Yi-Hsueh, L., Xiang, L., Bao, G., Dajiang, Z., Yixuan, Y., Dinggang, S.,

Tianming, L., and Shu, Z. (2023). Prompt Engineering for Healthcare: Methodologies and Applications. *arXiv.org*.

Jin, X., Vinzamuri, B., Venkatapathy, S., Ji, H., and Natarajan, P. (2023). Adversarial Robustness for Large Language NER models using Disentanglement and Word Attributions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Jung, V. and van der Plas, L. (2024). Understanding the effects of language-specific class imbalance in multilingual fine-tuning. *Findings*.

Kang, H., Ni, J., and Yao, H. (2023). Ever: Mitigating Hallucination in Large Language Models through Real-Time Verification and Rectification. *arXiv.org*.

Kochanek, M., Cichecki, I., Kaszyca, O., Szydło, D., Madej, M., Jędrzejewski, D., Kazienko, P., and Kocoń, J. (2024). Improving Training Dataset Balance with ChatGPT Prompt Engineering. *Electronics*, 13(12):2255.

Li, X., Wang, L., Dong, G., He, K., Zhao, J., Lei, H., Liu, J., and Xu, W. (2023). Generative Zero-Shot Prompt Learning for Cross-Domain Slot Filling with Inverse Prompting. *Annual Meeting of the Association for Computational Linguistics*.

Li, Z., Xu, X., Shen, T., Xu, C., Gu, J.-C., Lai, Y., Tao, C., and Ma, S. (2024). Leveraging Large Language Models for NLG Evaluation: Advances and Challenges. *arXiv.org*.

Lo, L. S. (2023). The Art and Science of Prompt Engineering: A New Literacy in the Information Age. *Internet Reference Services Quarterly*, 27(4):203–210.

Mansouri, A., Affendey, L., and Mamat, A. (2008). Named entity recognition approaches. *Int J Comp Sci Netw Sec*, 8.

Meskó, B. (2023). Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *Journal of Medical Internet Research*, 25:e50638.

Monajatipoor, M., Yang, J., Stremmel, J., Emami, M., Mohaghegh, F., Rouhsedaghat, M., and Chang, K.-W. (2024). Llms in Biomedicine: A study on clinical Named Entity Recognition. *arXiv.org*.

Park, Y.-J., Pillai, A., Deng, J., Guo, E., Gupta, M., Paget, M., and Naugler, C. (2024). Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Medical Informatics and Decision Making*, 24(1).

Rathod, J. D. (2024). Systematic Study of Prompt Engineering. *International Journal for Research in Applied Science and Engineering Technology*, 12(6):597–613.

Reynolds, L. and McDonell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7. ACM.

Russe, M. F., Reisert, M., Bamberg, F., and Rau, A. (2024). Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*.

Sellemann, B. (2021). Herausforderungen der Digitalisierung in der Pflege. *Public Health Forum*, 29(3):245–247.

Sonntagbauer, M., Haar, M., and Kluge, S. (2023). Künstliche Intelligenz: Wie werden ChatGPT und andere KI-Anwendungen unseren ärztlichen Alltag verändern? *Medizinische Klinik - Intensivmedizin und Notfallmedizin*, 118(5):366–371.

Treder, M. S., Lee, S., and Tsvetanov, K. A. (2024). Introduction to Large Language Models (LLMs) for dementia care and research. *Frontiers in Dementia*, 3.

Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., Yue, C., Zhang, H., Liu, Y., Pan, Y., Liu, Z., Sun, L., Li, X., Ge, B., Jiang, X., Zhu, D., Yuan, Y., Shen, D., Liu, T., and Zhang, S. (2023a). Prompt Engineering for Healthcare: Methodologies and Applications. *arXiv.org*.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., and Wang, G. (2023b). Gpt-NER: Named Entity Recognition via Large Language Models. *arXiv.org*.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv.org*.

Yu, J., Bohnet, B., and Poesio, M. (2020). Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Yuan, Y., Gao, J., and Zhang, Y. (2017). Supervised learning for robust term extraction. In *2017 International Conference on Asian Language Processing (IALP)*, volume 1031, pages 302–305. IEEE.

Zernikow, J., Grassow, L., Gröschel, J., Henrion, P., Wetzel, P. J., and Spethmann, S. (2023). Anwendung von "large language models" in der Klinik. *Die Innere Medizin*, 64(11):1058–1064.

Zhang, J., Li, Z., Das, K., Malin, B. A., and Kumar, S. (2023). Sac3: Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency. *Conference on Empirical Methods in Natural Language Processing*.

Zhou, C., He, J., Ma, X., Berg-Kirkpatrick, T., and Neubig, G. (2022). Prompt Consistency for Zero-Shot Task Generalization. *Conference on Empirical Methods in Natural Language Processing*.

Zhou, G. and Su, J. (2001). Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 473. Association for Computational Linguistics.