Resolving Elliptical Compounds in German Medical Text

Niklas Kämmer*, Florian Borchert*, Silvia Winkler,
Gerard de Melo, Matthieu-P. Schapranow
Hasso Plattner Institute, University of Potsdam, Germany
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam
`{firstname.lastname}@hpi.de`

Abstract

Elliptical coordinated compound noun phrases (ECCNPs), a special kind of coordination ellipsis, are a common phenomenon in German medical texts. As their presence is known to affect the performance in downstream tasks such as entity extraction and disambiguation, their resolution can be a useful preprocessing step in information extraction pipelines. In this work, we present a new comprehensive dataset of more than 4,000 manually annotated ECCNPs in German medical text, along with the respective ground truth resolutions. Based on this data, we propose a generative encoder–decoder Transformer model, allowing for a simple end-to-end resolution of ECCNPs from raw input strings with very high accuracy (90.5 % exact match score). We compare our approach to an elaborate rule-based baseline, which the generative model outperforms by a large margin. We further investigate different scenarios for prompting large language models (LLM) to resolve ECCNPs. In a zero-shot setting, performance is remarkably poor (21.6 % exact matches), as the LLM tends to apply complex changes to the inputs unrelated to our specific task. We also find no improvement over the generative model when using the LLM for post-filtering of generated candidate resolutions. The source code including instructions on how to access the data are available at: `https://github.com/hpi-dhc/ggponc_ellipses`.

1 Introduction

Coordination ellipses are a common phenomenon in many human languages, where certain words or phrases are omitted in a sentence due to their repetition from a previous or subsequent clause in a conjunction. Usually, we assume that the reader can infer the omitted words from the context. In the biomedical domain, the presence of coordination ellipses has been shown to negatively affect downstream tasks such as Named Entity Recognition (NER) (Buyko et al., 2007; Chae et al., 2014) and Entity Linking (Yuan et al., 2020).

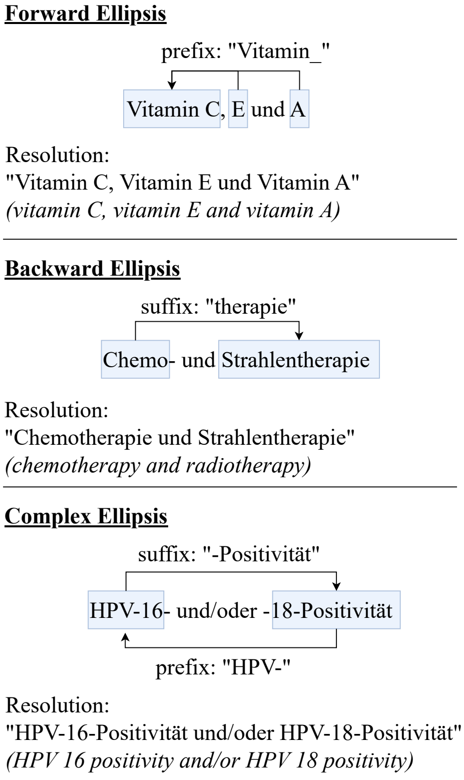
Forward Ellipsis
prefix: "Vitamin_"
Vitamin C, E und A
Resolution:
"Vitamin C, Vitamin E und Vitamin A"
(vitamin C, vitamin E and vitamin A)
Backward Ellipsis
suffix: "therapie"
Chemo- und Strahlentherapie
Resolution:
"Chemotherapie und Strahlentherapie"
(chemotherapy and radiotherapy)
Complex Ellipsis
suffix: "-Positivität"
HPV-16- und/oder -18-Positivität
prefix: "HPV-"
Resolution:
"HPV-16-Positivität und/oder HPV-18-Positivität"
(HPV 16 positivity and/or HPV 18 positivity)

Figure 1: Examples for ECCNPs, their annotation and resolution in GGPONC, a corpus of German oncology guidelines (Borchert et al., 2022)

In this work, we consider the special case of elliptical coordinated compound noun phrases (ECCNPs) in the German language, where they are commonly encountered in biomedical and other scientific texts, due to the high density of contained information and their compact form of writing. For instance, consider doctor's letters and medical reports from kidney transplant patients or cancer patients. Following the terminology introduced by Chae et al. (2014), we distinguish *forward*, *backward* and *complex ellipses*, as shown in Fig. 1. The

---

*Authors marked by * equally share first authorship.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 292–305
July 13, 2023 ©2023 Association for Computational Linguistics

resolution of ECCNPs is particularly challenging in German due to its closed noun compounds, i.e., noun compounds without separating space.

Recently, an increasing number of German medical corpora with semantic annotations have been released (cf. Zesch and Bewersdorff 2022 for a recent survey). The annotation schemes of some of these corpora, including BRONCO (Kittner et al., 2021) as well as the German subset of the Mantra GSC (Kors et al., 2015), support fragmented, i.e., non-contiguous, entity annotations. However, the annotations in these corpora do not reveal which of these are ECCNPs, as non-contiguous entity spans may also occur for various other reasons. At the same time, there may be further ECCNPs that are not among the annotated named entities. In the absence of a suitable dataset for training and evaluating ECCNP resolvers for German medical language, we provide a new, comprehensive dataset based on a manual re-annotation of a subset of GGPONC 2.0 (Borchert et al., 2022). Our annotated dataset, source code, and trained models are available online (HPI-DHC, 2023).

Key contributions of this work are:

- A novel dataset of more than 4,000 manually annotated ECCNPs in German-language oncology guidelines,
- A generative neural encoder–decoder model that resolves ECCNPs in German medical text in an end-to-end fashion, i.e., directly generating a string with completed elliptical compounds for a given input string,
- A comparison of our approach against prior work (Aepli and Volk, 2013) that adopted rule-based techniques, and
- An evaluation of the capabilities of the large language model (LLM) GPT-3.5 (OpenAI, 2023) on this task, both in a zero-shot manner and in different multiple choice settings.

The remainder of this work is structured as follows: in Section 2 we set our work in the context of related work. In Section 3 we present our novel dataset and the employed methods. We evaluate the performance of different approaches for resolving ECCNPs in Section 4. We discuss our findings in Section 5 and conclude our work with an outlook in Section 6.

2 Related Work

The identification and resolution of various kinds of coordination ellipses have been studied extensively. It is generally assumed that their presence leads to ambiguity, and some authors have even considered them as instances of *ill-formedness* (Lavelli and Stock, 1990). In the biomedical domain, their relevance was mostly discussed in the context of named entities and the resulting ambiguities in their detection and normalization.

Buyko et al. (2007) address the problem of resolving coordination ellipses inside named entities. Their approach identifies potential candidate spans through an NER step, followed by the identification of conjuncts, and a rule-based heuristic to resolve the ellipses. For identifying conjuncts, the authors find that a Conditional Random Field (CRF) outperforms a rule-based baseline. Buyko and Hahn (2008) evaluate different types of features for the conjunct identification step. Similarly, Wei et al. (2015) use a CRF with a variety of morphosyntactic and semantic features for recognizing coordination ellipses, followed by a rule-based approach for reassembling the segmented tokens. Chae et al. (2014) provide an extensive formal analysis of various types of omissions in coordinated noun phrases and propose a rule-based algorithm, also relying on the identification of named entities (here with a dictionary). Blake and Rindflesch (2017) consider the special case of ECCNPs and use a dependency parser for identifying candidate terms, and then generate potential resolutions using syntactical and semantic information. Yuan et al. (2020) propose RECEEM, an unsupervised, graph-based approach for reconstructing elliptical named entities in previously identified entity spans. Importantly, they show the positive effect of resolving such ellipses for downstream Entity Linking performance.

All the aforementioned approaches build on rather complex multi-stage pipelines with a range of assumptions. Most of them consider ellipses inside named entities only and rely on NER to identify candidate spans. They also employ different heuristics, usually based on syntactic information provided through part-of-speech (POS) tagging and dependency parsing. Furthermore, most of the existing approaches assume that conjuncts are given as individual tokens, with Wei et al. (2015) being a notable exception. This might be a sensible assumption for the English language, but it makes it difficult to apply these approaches to languages where compound nouns are usually not tokenized (such as German). To the best of our knowledge, Aepli and Volk (2013) have proposed the only algorithm that deals with ECCNPs for German text, albeit not in the biomedical domain. We use this work as a baseline to evaluate our approach and thus describe it in further detail in Section 3.2.

In contrast to prior work, our proposed generative approach treats the problem in an end-to-end fashion by directly resolving all ECCNPs in an input sentence. Therefore, it does not rely on any particular tokenization or segmentation of the input. We note that Quan et al. (2019) have proposed a similar generative approach, although they consider different types of ellipses in task-oriented dialogue and rely on dialogue context.

3 Materials and Methods

In the following, we describe our contributed dataset, different methods for resolving ECCNPs, and metrics for the evaluation of such methods.

3.1 Dataset and Annotation

Our dataset is based on entity annotations in GGPONC 2.0, which were created by a team of seven medical students (see Borchert et al. 2022 for details). As an additional annotation layer in the original project, annotators used *fragment* relations to connect elliptical conjuncts with their headword, as exemplified in Fig. 1. Although the reported inter-annotator agreement for GGPONC 2.0 is very high (γ=.94), we observed some inconsistencies in the annotations required to resolve ECCNPs. For instance, suspensive hyphens were occasionally but not always among the annotated tokens.

To address these inconsistencies, we employed two human annotators to review and correct all documents (i.e., guideline segments) containing at least one ECCNP. One of them was a medical student with substantial experience in linguistic annotation and the other a computational linguist. In total, 2,601 such documents were reviewed and used in this work, which is around a quarter of the original corpus. Although the original annotations cover other types of elliptical coordinated noun phrases, our study only considered ellipses occurring inside compound noun phrases. The annotators added instances of ECCNP annotations where they were missing, e.g., when the spans were not covered by the original entity definitions of GGPONC 2.0.

The statistics of our dataset are given in Table 1.

	Train.	**Dev.**	**Test**	**Total**
Sentences	5,336	1,201	1,159	7,696
w/ Ellipses	2,715	590	543	3,848
Controls	2,621	611	616	3,848
Ellipses	2,896	629	586	4,111
Forward	512	101	101	714
Backward	2,271	499	458	3,228
Complex	113	29	27	169

Table 1: Descriptive statistics of our new dataset of ECCNPs and their resolutions. We report the number of sentences and ellipses in these sentences in the **Train**ing, **Dev**elopment and **Test** split as well as total counts.

In total, we annotated 4,111 instances of ECCNPs in 3,848 sentences (241 sentences contain two or more ECCNPs). 714 of these were forward, 3,228 backward, and 169 were complex ellipses. The set of documents was split into 70 % training, 15 % development, and 15 % test data for model training and evaluation. The splits were assigned at the document rather than the sentence level to prevent leakage from repeated occurrences of ellipses within the same documents. As controls, we sampled the same number of sentences (3,848) not containing any ECCNP from the set of manually reviewed documents.

3.2 Rule-based Baseline

As a baseline, we adopt the rule-based approach proposed by Aepli and Volk (2013), who devised a two-step algorithm that first identifies elliptical compounds in a given sentence and subsequently resolves them in a second step. Their algorithm requires several inputs: the tokenized and POS-tagged text in a custom XML format, a list of word frequencies, and pre-computed word segmentations. In the original paper, a corpus with manual POS annotations was used. Since GGPONC does not provide such annotations, we used the SPACY library with the `de_core_news_lg` model for tokenization and POS tagging (Montani et al., 2022). Due to the unavailability of the original morphological analyzer GERTWOL, we instead used CHARSPLIT, an n-gram-based compound splitter as described by Tuggener (2016). Word frequencies are derived from the 1.87M tokens in the complete GGPONC corpus.

In the first step, the algorithm described by Aepli and Volk (2013) detects coordination ellipses based on pre-defined patterns using POS tags. For instance, the backward ellipsis "Detektions-

und Inzidenzraten” (detection [rates] and incidence rates) would match a pattern for a truncated word (Trunc) followed by a conjunction (Konj) and a noun (NN). For each identified pattern, the complete compound is extracted and split using precomputed word segmentations. For the aforementioned example, the model would split “Inzidenzraten” into “Inzidenz” and “raten”. Lastly, the lexeme “raten” would be used to complete the elliptical conjunct “Detektions-” to "Detektionsraten" and replace it in the input sentence. If a word cannot be segmented by the morphological analyzer, or when solutions are ambiguous, word frequencies are used to select the most probable segmentation.

3.3 Generative Encoder–Decoder Model

For our proposed approach, we use a Transformer-based encoder–decoder architecture for text generation. Specifically, we use the base version of mT5, the multilingual version of the T5 model (Xue et al., 2021; Raffel et al., 2020). mT5 closely follows the original T5 architecture, but was trained on a multilingual Common Crawl-based dataset called *mC4* with 101 languages, including German.

We performed some preliminary experiments with other models. These included a German T5 model, a regular and a German byte-oriented ByT5 model, as well as M2M100 and mBART (Einmalumdiewelt, 2022; Xue et al., 2022; A-Ware, 2021; Fan et al., 2021; Liu et al., 2020). All of them showed inferior performance to mT5.

The mT5 model is trained with complete input and output sentences from the training dataset. For hyperparameter optimization, we use Bayesian hyperparameter search as implemented in the Optuna framework (Akiba et al., 2019). The optimized hyperparameters are given in Table 2. All models were trained for 20 epochs on a single Nvidia A40 GPU. For the final model, we chose the configuration that maximized exact matches with the ground truth in the development set. The optimal parameters as well as the hyperparameter search settings are made available as a Hydra configuration file (Yadan, 2019) together with our source code repository (HPI-DHC, 2023).

At inference time, we retrieve the top k most likely generations using beam search (Freitag and Al-Onaizan, 2017), i.e., we greedily traverse the search tree as the model generates the output from left to right and maintains a fixed amount of the most probable candidate sentences.

Learning Rate	$[1 \times 10^{-6}, 1 \times 10^{-3}]$
Batch Size	4, 8, 16
Weight Decay	[0, 1]
Warmup Steps	[0, 500]
Max. Generation Length	[50, 500]

Table 2: Optimized hyperparameters for the generative model and their ranges

3.4 Prompting LLMs for Zero-Shot Resolution

Recent advances in LLM capabilities motivated us to further evaluate the zero-shot performance of GPT-3.5 (OpenAI, 2023) on our task. Using a subset of the development set, we created several prompts to explain the task with varying amounts of context and examples. The best-performing prompt (see Appendix A.1) was used to perform a full evaluation on the test set.

For all experiments, we use GPT-3.5 (`gpt-3.5-turbo-0301`) through the paid OpenAI API, setting the `temperature` parameter to 0 and restricting `max_tokens` to 400. We also repeated some of our initial experiments on the instruction model `text-davinci-003`, which did not yield improved performance. In total, we spent slightly less than $50 on OpenAI API usage, which includes preliminary experiments with the more expensive `text-davinci-003` model and several iterations of sending individual prompts for each sentence in the entire validation and test sets.

3.5 Post-Filtering of Top k Generated Outputs

In addition to the zero-shot scenario, we also evaluate the ability of GPT-3.5 to choose from a given list of possible resolutions. This could be particularly useful when the top 1 result of our model is wrong, but the correct answer is among the top k retrieved results.

We evaluate two approaches to choose from the top k outputs of our encoder–decoder model generated through beam search:

1. A multiple-choice prompt, which explains the problem of resolving ECCNPs using an example, lists the top k generated outputs and asks the LLM to return the index of the correct resolution (see Appendix A.2)

2. A two-step approach that emphasizes the top 1 generation to the LLM by first asking the model to evaluate whether the given output is

the correct resolution (which it likely is). The LLM is asked to return either “Yes” or “No”. If the answer is positive, we keep the top 1 answer. Otherwise, we prompt the model with the remaining $k-1$ generations along with a request to return the index of the correct resolution (see Appendix A.3).

As an additional heuristic, we assume that no ECCNPs are present in sentences where the top 1 prediction of the encoder–decoder model is equal to the input. While this could potentially lead to a few additional false negatives, it reduces the number of API calls by around 50 %. For all experiments, we set $k=5$ unless stated otherwise, which is again a trade-off to balance the recall of correct answers and the (token-based) cost of API calls.

3.6 Evaluation Metrics

We consider different metrics for evaluating our model predictions. A straightforward metric is the *exact match* of generated output compared to the gold standard. However, this metric might be overly strict when a generative model makes minor errors that do not alter the meaning of the sentence (e.g., missing or different punctuation), or in cases where the actual resolution is slightly ambiguous. At the same time, standard metrics for evaluating generative models, such as GLEU (Wu et al., 2016), will be trivially high even when the model fails entirely, as the desired output typically differs from the input by just a few characters.

Due to these shortcomings of traditional metrics, we propose a score based on the *relative edit distance* of generated output, gold standard, and input. Let Δ_{edit} be the unnormalized Levenshtein distance, p the predicted string, g the ground truth, and i the input string. Then, the relative edit distance score is defined as $s_{\text{edit}}(p, g; i) = 1$ if $\Delta_{\text{edit}}(p, g) = 0$ (true positive or true negative) and otherwise:

$$s_{\text{edit}}(p, g; i) = 1 - \frac{\Delta_{\text{edit}}(p, g)}{\Delta_{\text{edit}}(p, i) + \Delta_{\text{edit}}(i, g)}$$

This score has the property of being:

- 1 for exact matches,
- 0 for false negatives and false positives, i.e., either $\Delta_{\text{edit}}(p, i)$ or $\Delta_{\text{edit}}(i, g)$ are 0,
- between 0 and 1 otherwise.

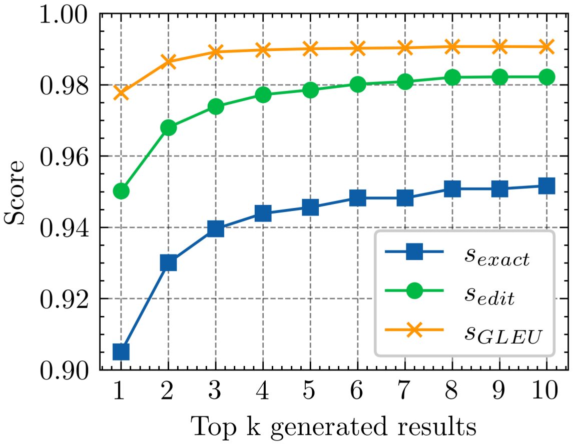

Figure 2: Top k results for increasing numbers k of candidates generated through beam search from our encoder–decoder model

For all experiments, we report the exact match (s_{exact}), GLEU score (s_{GLEU}), and the proposed relative edit distance (s_{edit}).

3.7 Error Types

To characterize situations where $s_{\text{edit}}(p, g; i)$ is neither 0 nor 1, we categorize errors based on the operations that would turn p into g, i.e., by the relative errors introduced in the prediction. We distinguish *deletions*, *insertions*, *replacements*, as well as *complex* errors, where the latter indicates any combination of the three former kinds of errors in the same sentence. To determine such operations, we invoke the `difflib` module in the PYTHON 3 standard library.

4 Results and Evaluation

Table 3 summarizes performance metrics and error counts on the test set from different experiments.

4.1 Model Performance

Our supervised encoder–decoder model noticeably outperforms all other approaches. However, the rule-based baseline is more conservative and yields only a single false positive generation for the control dataset without ellipses. The zero-shot results of GPT-3.5 substantially fall behind both our supervised approach and the rule-based baseline.

Fig. 2 plots the results of our generative model when considering the best resolution among k candidates generated through beam search. All scores, including s_{exact}, improve with increasing k, showing that the correct resolution is regularly among the top k candidates even if it is not the most likely

	Scores			Matches		Errors					
	s_{exact}	s_{edit}	s_{GLEU}	TP	TN	FP	FN	Ins.	Del.	Rep.	Com.
Rule-based Baseline	0.701	0.771	0.965	198	**615**	1	227	11	36	50	21
Encoder–Decoder (mT5)	**0.905**	**0.950**	0.978	**448**	601	15	18	25	25	16	11
GPT 3.5 (Zero-Shot)	0.216	0.257	0.830	52	198	418	100	8	45	72	266
GPT 3.5 (Multiple Choice)	0.732	0.912	0.969	238	610	6	54	83	71	80	17
GPT 3.5 (Two-Step)	0.897	0.948	**0.981**	438	602	14	19	26	27	22	11

Table 3: Results on the held-out test set (543 sentences with ellipses, 616 controls). We report aggregate scores and a breakdown of exact matches (**T**rue **P**ositives and **T**rue **N**egatives) as well as different error conditions (**F**alse **P**ositives, **F**alse **N**egatives, **Ins**ertions, **Del**etions, **Rep**etitions, and **Com**plex Errors. Highest values for each score and exact matches are highlighted in **bold**.

	Example 1	
Input	Wirts- [REF] [REF] [REF], Umwelt- [REF] und bakteriellen Virulenzfaktoren [REF] *(Host [REF] [REF] [REF], environmental [REF] and bacterial virulence factors [REF].)*	
Encoder–Decoder	Wirts**faktoren** [REF] [REF] [REF], Umwelt**faktoren** [REF] und bakteriellen Virulenzfaktoren [REF]	TP ✓
Zero-Shot	Wirts-**faktoren**, Umwelt-**faktoren** und bakteriellen Virulenzfaktoren	Com. ✗
	Example 2	
Input	83,3% in ersterer Gruppe und 82,7% in letzterer *(83.3% in the former group and 82.7% in the latter.)*	
Encoder–Decoder	83,3% in ersterer Gruppe und 82,7% in letzterer	TP ✓
Zero-Shot	83,3% in der ersten Gruppe und 82,7% in der letzten Gruppe	FP ✗

Table 4: Examples for complex errors and false positives introduced by the zero-shot approach. We highlight the omitted prefixes and suffixes in **bold** and underline incorrectly inserted characters. In Ex. 1, hyphenations are not replaced and literature references, unrelated to the task, are removed. Although no ellipses are present in Ex. 2, the model rephrases the sentence and simultaneously resolves an unrelated kind of ellipsis.

output according to the decoder. However, providing these options as examples to choose from to GPT-3.5 is generally inferior to selecting the most likely candidate, apart from a negligible improvement in GLEU score for the two-step approach.

4.2 Error Analysis

The GPT-3.5 zero-shot approach often introduces substantial undesired changes to the sentence structure: frequently, ECCNPs are indeed resolved, but various other changes are encountered in the output (see Table 4). However, we noted that many changes introduced by GPT-3.5 are unrelated to the task while keeping the overall semantics of the input sentence.

The rule-based baseline and generative model both reliably detect and resolve simple backward ellipses with suspensive hyphens and common conjunctions, e.g., “und” (and) or “oder” (or), as shown in Table 5, Ex. 3. However, the rule-based baseline fails to detect many ECCNPs with other commonly used expressions or prepositions, e.g., “/” (slash), “bzw.” (respectively), “vs.”, or “versus” as shown in Table 5, Ex. 4.

The baseline is generally unreliable for long-range dependencies (Table 5, Ex. 5) and forward ellipses (Table 5, Ex. 6). Indeed, ten out of eleven rules implemented by Aepli and Volk (2013) concern backward ellipses and rely on the presence of the part-of-speech tag TRUNC, which is not reliably detected by SPACY in expressions such as “BRAF- und MEK-Mutation” (BRAF and MEK mutation). It also does not apply to multi-token noun phrases such as “Vitamin A” in “Vitamin A und B” (Vitamins A and B). Generally, the baseline produces almost no false positives, but many *replacement* errors. This is particularly true for complex ellipses (Table 5, Ex. 7), where multiple rules apply simultaneously.

	Example 3	
Input	chronischer Tabak- oder Alkoholabusus *(Chronic tobacco or alcohol abuse)*	
Ground Truth	chronischer Tabak**abusus** oder Alkoholabusus	
Rule-Based	chronischer Tabak**abusus** oder Alkoholabusus	TP ✓
Encoder–Decoder	chronischer Tabak**abusus** oder Alkoholabusus	TP ✓
	Example 4	
Input	Zangen- versus Schlingenabtragung kleiner Polypen *(Forceps versus snare ablation of small polyps)*	
Ground Truth	Zangen**abtragung** versus Schlingenabtragung kleiner Polypen	
Rule-Based	Zangen- versus Schlingenabtragung kleiner Polypen	FN ✗
Encoder–Decoder	Zangen**abtragung** versus Schlingenabtragung kleiner Polypen	TP ✓
	Example 5	
Input	Plattenepithel- sowie für die Mehrzahl der Adenokarzinome *(Squamous cell carcinomas and the majority of adenocarcinomas)*	
Ground Truth	Plattenepithel**karzinome** sowie für die Mehrzahl der Adenokarzinome	
Rule-Based	Plattenepithelzahl sowie für die Mehrzahl der Adenokarzinome	Repl. ✗
Encoder–Decoder	Plattenepithel**karzinome** sowie für die Mehrzahl der Adenokarzinome	TP ✓
	Example 6	
Input	Stomaprolaps, Stomaretraktion, -stenose *(Stoma prolapse, stoma retraction, stoma stenosis)*	
Ground Truth	Stomaprolaps, Stomaretraktion, **Stoma**stenose	
Rule-Based	Stomaprolaps, Stomaretraktion, -stenose	FN ✗
Encoder–Decoder	Stomaprolaps, Stomaretraktion, **Stoma**restenose	Ins. ✗
Best of Top 5	Stomaprolaps, Stomaretraktion, **Stoma**tstenose	Ins. ✗
	Example 7	
Input	Stadium IIA- und -IIB-Seminom *(Stage IIA and IIB seminoma)*	
Ground Truth	Stadium IIA-**Seminom** und **Stadium** IIB-Seminom	
Rule-Based	Stadium IIA-s**eminom** und IiaIIBIiaSeminom	Repl. ✗
Encoder–Decoder	Stadium IIA-**Seminom** und **Stadium** IIB-Seminom	TP ✓
	Example 8	
Input	Die bildgebenden Verfahren CT, MRT oder FDG-PET-CT *(The imaging techniques CT, MRI or FDG-PET-CT)*	
Ground Truth	Die bildgebenden Verfahren CT, MRT oder FDG-PET-CT	
Rule-Based	Die bildgebenden Verfahren CT, MRT oder FDG-PET-CT	TN ✓
Encoder–Decoder	Die bildgebenden Verfahren CT-CT, MRT-CT oder FDG-PET-CT	FP ✗
Best of Top 5	Die bildgebenden Verfahren CT, MRT oder FDG-PET-CT	TP ✓

Table 5: Examples of ECCNPs and their resolutions by the rule-based baseline (Aepli and Volk, 2013) and the generative encoder–decoder model with the respective error types. When the generative model's top 1 prediction is incorrect (Ex. 6 and 8), we also provide the best answer from the top 5 outputs generated through beam search.

In contrast to the baseline, the generative encoder–decoder model does occasionally produce false positives, in particular in the presence of abbreviations, non-suspensive hyphens, and minus signs (Table 5, Ex. 8). However, it is usually robust when no hyphen is present (as in the case of "Vitamin A und B"). In many cases, an exact match or more adequate resolution is among the top k candidates, as seen in Table 5, Ex. 6 and 8. The former example illustrates a common error scenario, where the generated output differs from the ground truth by very few characters, which hardly affects s_{edit} and s_{GLEU}.

5 Discussion

In the following, we further discuss our findings and potential alternative experimental designs.

5.1 Prompt Engineering

Deriving a prompt that would yield the best possible performance from a model like GPT-3.5 posed a substantial challenge (Reynolds and McDonell, 2021; Jimenez Gutierrez et al., 2022; Madaan et al., 2022). While the LLM appears to be able to follow the basic instructions for the task rather easily (independently of how the problem is framed), small changes in the prompt can drastically affect the performance when evaluated on the entire test set. This is particularly apparent in the zero-shot setting, where the in-context examples shown to the model have a substantial influence.

Our prompt engineering process was highly experimental. We evaluated our prompts on a comparable small subset of samples and chose the best-performing ones with minor variations. Furthermore, only a few prompts were evaluated on the entire dataset. As a result, we believe that there is room for improvement by using more elaborate approaches for prompt engineering.

5.2 Different Task Formulations

We notice that different translations of the task into LLM prompts led to a large variability in performance. Transforming the task into multiple-choice questions yielded sensible results, despite not improving over simply choosing the first answer delivered by the generative model. Simplifying the task by highlighting the most likely answer in a two-step approach increases the performance. It may be worthwhile to apply more fine-grained post-filtering operations and subdivide the task by asking the LLM for a binary decision on each generated output (an extension of the two-step approach). However, the basis for this approach would still be a well-performing generative model. It would also increase the complexity of a simple end-to-end model substantially. Overall, the results of our attempts to integrate GPT-3.5 into the task have to be considered preliminary. However, the low zero-shot performance relative to carefully tuned supervised models is not unique to our problem (Jimenez Gutierrez et al., 2022; Yuan et al., 2023; Bang et al., 2023; Hu et al., 2023), in particular for languages other than English (Lai et al., 2023).

5.3 Impact on Downstream Tasks

Frequently, both the encoder–decoder and the zero-shot approach fail in subtle ways. Often only a single character or punctuation mark is incorrect in the predicted string (e.g., see Ex. 6 in Table 5). In our evaluation, this is counted as a complete mismatch in the exact matching score. While the zero-shot approach often fails spectacularly, it sometimes merely removes redundant information (e.g., inline literature references), performs additional tasks (such as abbreviation expansion), or applies semantically equivalent rewritings of the text, as shown in Table 4. While all of these operations are at odds with the specific task at hand and merit a reduction of the evaluation scores, they may be irrelevant for some downstream tasks, such as semantic indexing with document-level predictions.

6 Conclusion and Future Work

In this work, we have presented a novel dataset and an end-to-end approach for the task of resolving ECCNPs in German medical text. A supervised encoder–decoder approach yields strong results on this task, substantially outperforming prior work, which was rule-based. While the general-purpose model mT5 worked well in our initial experiments, a systematic evaluation and hyperparameter optimization for other architectures and checkpoints from the increasingly large pool of text generation models may yield further improvements, e.g. the evaluation of larger versions of mT5 (Large, XL, or XXL) with our pipeline.

Although the initial zero-shot results with GPT-3.5 were rather disappointing, they could potentially be improved using more capable LLMs – at the time of this writing, access to the most recent release by OpenAI (GPT-4) was still in a restricted rollout phase with a waiting list. Such models might also benefit from fine-tuning on labeled examples. Further investigation of the root causes for low exact match scores might also help in designing more appropriate prompts and choosing examples for in-context learning. Additionally, we plan to investigate different ways to choose from the top k generated outputs of our encoder–decoder model, e.g., by framing it as a supervised multiple-choice or natural language inference problem.

In an earlier study, we have shown that a modern, BERT-based NER model was remarkably robust in the presence of coordination ellipses (Borchert et al., 2022). However, other prior work suggests that the resolution of ECCNPs would still benefit downstream tasks like entity linking. We plan to systematically assess the impact of resolving ECCNPs in German text for such tasks in future work to determine the broad applicability of our approach. Considering the simplicity of the encoder–decoder model, it can be easily integrated into clinical NLP pipelines, which are increasingly becoming available for the German language and medical domains, in particular oncology, cardiology, and nephrology.

Limitations

One limitation of this study is a potential lack of reproducibility for some experiments due to the employment of the proprietary GPT 3.5 LLM. In the future, this could be addressed by deploying and versioning open-source LLMs with similar capacities locally. We acknowledge a potential bias in our dataset, as only a quarter of the full corpus was manually reviewed based on existing annotations, i.e., three-quarters of GGPONC 2.0 were excluded from our analysis so far. Therefore, the matched dataset of controls will likely overrepresent the prevalence of ECCNPs, while not adequately representing the full spectrum of hard negative cases. However, we considered a manual review of the full corpus as unnecessary due to the very low number of ECCNPs expected in the remaining documents. Further, we note that also other kinds of coordination ellipses (omitted adjectives or verbs) besides ECCNPs may be relevant for downstream information extraction tasks. Although preliminary annotations for these exist in GGPONC 2.0, we excluded them from our analysis, as annotations were not comprehensive, and their resolution is often more ambiguous than elliptical compounds.

Ethics Statement

We have used clinical guidelines as a foundation for this study because they are not related to individual patients and preclude any potential risk of harm or violation of privacy. They do not contain any personal health information (PHI) or personally identifiable information (PII) whilst containing specific German biomedical terms. We acknowledge that reproducing the experiments using other genres of clinical text from a patient context may pose ethical challenges. For example, consider the medical reports contained in the history of kidney transplant patients or cancer patients. Such medical reports contain very sensitive information over months and years of medical treatment of individual patients, but also expose information relevant to family members at risk. Medical reports from historic patient cases might be of great benefit for developing clinical prediction models. However, sending such sensitive data containing PHI or PII to a third-party service provider (like OpenAI) might lead to the exposure of personal data, privacy breaches, and ethical protocol violations. As a result, we recommend the use of open-source LLMs, which can also be deployed on-premise. Such solutions mitigate these risks and address ethical concerns, but also enhance reproducibility.

Acknowledgements

Parts of this work were generously supported by grants of the German Federal Ministry of Research and Education (01ZZ1802H, 01ZZ2314N) and the German Federal Ministry of Economic Affairs and Climate Action (01MJ21002A).

References

A-Ware. 2021. aware-ai/byt5-german-grammar. `https://huggingface.co/aware-ai/byt5-german-grammar` [retrieved: Apr 18, 2023].

Noëmi Aepli and Martin Volk. 2013. Reconstructing complete lemmas for incomplete German compounds. In *Language Processing and Knowledge in the Web*, pages 1–13. Springer.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Catherine L. Blake and Thomas C. Rindflesch. 2017. Leveraging syntax to better capture the semantics of elliptical coordinated compound noun phrases. *Journal of Biomedical Informatics*, 72:120–131.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P Schapranow. 2022. GGPONC 2.0 - the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3650–3660, Marseille, France. European Language Resources Association.

Ekaterina Buyko and Udo Hahn. 2008. Are morpho-syntactic features more predictive for the resolution of noun phrase coordination ambiguity than lexico-semantic similarity scores? In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 89–96, Manchester, UK.

Ekaterina Buyko, Katrin Tomanek, and Udo Hahn. 2007. Resolution of coordination ellipses in biological named entities using conditional random fields. In *PACLING 2007 — Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 163–171. Pacific Association for Computational Linguistics.

Jeongmin Chae, Younghee Jung, Taemin Lee, Soonyoung Jung, Chan Huh, Gilhan Kim, Hyeoncheol Kim, and Heungbum Oh. 2014. Identifying non-elliptical entity mentions in a coordinated NP with ellipses. *Journal of Biomedical Informatics*, 47:139–152.

Einmalumdiewelt. 2022. Einmalumdiewelt/t5-base_gnad. `https://huggingface.co/Einmalumdiewelt/T5-Base_GNAD` [retrieved: Apr 18, 2023].

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

HPI-DHC. 2023. GGPONC ellipses source code repository. `https://github.com/hpi-dhc/ggponc_ellipses` [retrieved: Apr 18, 2023].

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, et al. 2021. Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA open*, 4(2):ooab025.

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M Van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Alberto Lavelli and Oliviero Stock. 1990. When something is missing: Ellipsis, coordination and the chart. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*, pages 184–189.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve GPT-3 after deployment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2833–2861, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ines Montani, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, Jim Geovedi, Jim O'Regan, Maxim Samsonov, Duygu Altinok, György Orosz, Daniël de Kok, Søren Lind Kristiansen, Lj Miranda, Explosion Bot, Roman, Peter Baumgartner, Leander Fiedler, Richard Hudson, Madeesh Kannan, Edward, Grégory Howard, Wannaphong Phatthiyaphaibun, Yohei Tamura, Sam Bozek, murat, Ryn Daniels, and Flusskind. 2022. explosion/SPACY: v3.4.1: Fix compatibility with CuPy v9.x.

OpenAI. 2023. GPT-3.5. `https://platform.openai.com/docs/models/gpt-3-5` [retrieved: Apr 18, 2023].

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. GECOR: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4547–4557, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Laria Reynolds and Kyle McDonell. 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.

Don Tuggener. 2016. *Incremental coreference resolution for German*. Ph.D. thesis, University of Zurich.

Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. 2015. SIMCONCEPT: a hybrid approach for simplifying composite named entities in biomedical text. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1385–1391.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Omry Yadan. 2019. Hydra - a framework for elegantly configuring complex applications. GitHub. `https://github.com/facebookresearch/hydra` [retrieved: Apr 18, 2023].

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt. *arXiv preprint arXiv:2304.05454*.

Chi Yuan, Yongli Wang, Ning Shang, Ziran Li, Ruxin Zhao, and Chunhua Weng. 2020. A graph-based method for reconstructing entities from coordination ellipsis in medical text. *Journal of the American Medical Informatics Association*, 27(9):1364–1373.

Torsten Zesch and Jeanette Bewersdorff. 2022. German medical natural language processing–a data-centric survey. In *The Upper-Rhine Artificial Intelligence Symposium UR-AI 2022 : AI Applications in Medicine and Manufacturing, 19 October 2022, Villingen-Schwenningen, Germany*, pages 137–145. Furtwangen University.

A GPT-3.5 Prompts

In the following, we provide examples to illustrate the prompts used in our experiments with GPT-3.5. We use the following sentence from our dataset as a running example:

Input: Bei HIV-/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektio- und/oder Gastroenterologen erfolgen.

Ground Truth: Bei HIV-**Positivität**/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektio**logen** und/oder Gastroenterologen erfolgen.

English translation: *In the case of HIV* ***positivity****/HCV positivity, further diagnostic and therapeutic management should take place at a specialised centre in interdisciplinary cooperation with infecti****ologists*** *and/or gastroenterologists.*

A.1 Zero-Shot

Prompt: »Ich werde dir im Folgenden einen Satz zeigen, welcher sogenannte Koordinationsellipsen enthält. Du sollst sie finden und auflösen.

Ein Beispiel: 'Dieser Unterschied zeigte sich sowohl im Stadium I/II als auch im Stadium III [REF].'

Deine Antwort: 'Dieser Unterschied zeigte sich sowohl im Stadium I/Stadium II als auch im Stadium III [REF].'.

Ein weiteres Beispiel: 'Ziel ist eine möglichst genaue Beschreibung der Tumormerkmale und –ausdehnung um eine optimale an das Erkrankungsstadium angepasste Therapie zu ermöglichen.'

Deine Antwort: 'Ziel ist eine möglichst genaue Beschreibung der Tumormerkmale und Tumorausdehnung um eine optimale an das Erkrankungsstadium angepasste Therapie zu ermöglichen.'

Ein weiteres Beispiel: 'Eine Metaanalyse von 16 randomisierten Studien mit BRAF- oder/und

MEK-Inhibitoren bestätigt die Überlegenheit der Kombinationstherapie [REF].'

Deine Antwort: 'Eine Metaanalyse von 16 randomisierten Studien mit BRAF-Inhibtoren oder/und MEK-Inhibitoren bestätigt die Überlegenheit der Kombinationstherapie [REF].'

Ein letztes Beispiel: 'Das Erkrankungs- und Sterberisiko ist jedoch deutlich unterschiedlich.'

Deine Antwort: 'Das Erkrankungsrisiko und Sterberisiko ist jedoch deutlich unterschiedlich.' Wenn ich dir gleich mein Beispiel gebe, antworte nur mit dem korrigierten Satz und keiner Erklärung.

Korrigiere bitte folgenden Satz: 'Bei HIV-/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektio- und/oder Gastroenterologen erfolgen.' «

Result: »Bei HIV- und HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektiologen und/oder Gastroenterologen erfolgen.«

English Translation of Prompt Template: *»I will show you a sentence that contains so-called coordination ellipses. You are supposed to find them and resolve them.*

Example: [...]

Your answer: [...]

Another example: [...]

Your answer: [...]

Another example: [...]

Your answer: [...]

One last example: [...]

Your answer: [...]

When I provide you with my example, please only respond with the corrected sentence without any explanation.

Please correct the following sentence: [...] «

A.2 Multiple Choice

Prompt: »Dir wird im Folgenden ein Satz gezeigt, welcher sogenannte Koordinationsellipsen enthält. Das Ziel ist es diese zu aufzulösen.

Ein Beispiel wäre 'Ibrutinib, ein Inhibitor der Bruton-Tyrosinkinase (BTK), ist in Deutschland als Erstlinien- und Rezidivtherapie in der CLL zugelassen.'

Die richtige Auflösung wäre 'Ibrutinib, ein Inhibitor der Bruton-Tyrosinkinase (BTK), ist in Deutschland als Erstlinientherapie und Rezidivtherapie in der CLL zugelassen.'

Dir werden zu den Beispielen Antwortmöglichkeiten gegeben und du sollst dann entscheiden, welche dieser Optionen die Koordinationsellipsen korrekt auflöst.

Der originale Satz: 'Bei HIV-/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektio- und/oder Gastroenterologen erfolgen.'

Deine Antwortmöglichkeiten:
1) 'Bei HIV-Positivität/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektioologen und/oder Gastroenterologen erfolgen.'
2) 'Bei HIV-Positivität/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektiologen und/oder Gastroenterologen erfolgen.'
3) 'Bei HIV-Positivität/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektiologos und/oder Gastroenterologen erfolgen.'
4) 'Bei HIV-/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektioologen und/oder Gastroenterologen erfolgen.'
5) 'Bei HIV-positivität/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektioologen und/oder Gastroenterologen erfolgen.'

Welche Antwort ist die richtige? Antworte nur mit der Zahl und keiner Erklärung «

Answer: »2«

Prediction: Bei HIV-Positivität/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektiologen und/oder Gastroenterologen erfolgen.

English Translation of Prompt Template: *»You will be shown a sentence that contains so-called* *coordination ellipses. The goal is to resolve them.*

An example would be: [...] The correct resolution would be: [...]

For the examples given, you will be provided with response options, and you should then decide which of these options correctly resolves the coordination ellipses.

The original sentence: [...]

Your response options:
1) [...]
2) [...]
3) [...]
4) [...]
5) [...] «

A.3 Two-Step

A.3.1 Example 1

Prompt 1: »Sie haben ein Modell entwickelt, das Koordinationsellipsen in Sätzen erkennt und auflöst. Das Modell gibt fünf verschiedene Versionen des ursprünglichen Satzes zurück, wobei die erste Version die wahrscheinlichste ist. Bitte lesen Sie sich die erste Version des Satzes sorgfältig durch und entscheiden Sie, ob diese Version korrekt ist und den ursprünglichen Satz mit aufgelösten Koordinationsellipsen wiedergibt. Bitte antworten Sie mir nur 'Ja' oder 'Nein' und keiner Erklärung!

Ursprünglicher Satz: 'Bei HIV-/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektio- und/oder Gastroenterologen erfolgen.'

Erste Version: 'Bei HIV-Positivität/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektioologen und/oder Gastroenterologen erfolgen.'«

Answer: »Ja.«

Result: Bei HIV-Positivität/HCV-Positivität sollte das weitere diagnostische und therapeutische Management an einem spezialisierten Zentrum in interdisziplinärer Zusammenarbeit mit Infektioologen und/oder Gastroenterologen erfolgen.

English Translation of Prompt Template: *»You have developed a model that recognizes and resolves coordination ellipses in sentences. The model provides five different versions of the original sentence, with the first version being the most probable. Please carefully read the first version of the sentence and decide whether this version is correct and represents the original sentence with resolved coordination ellipses. Please respond with either 'Yes' or 'No' without any explanation.*

Original sentence: [...]

First version: [...]«

A.3.2 Example 2

In the following example, the two-step approach proceeds to the second step by not accepting the suggested input.

Input: Die Tumoren werden mittels Kryosonden auf -60 °C bis -70 °C gekühlt, wodurch über die Herausbildung einer Koagulationsnekrose und Apoptose eine komplette Zell-, Gewebe- und damit Tumorzerstörung erreicht werden soll.

Ground Truth: Die Tumoren werden mittels Kryosonden auf -60 °C bis -70 °C gekühlt, wodurch über die Herausbildung einer Koagulationsnekrose und Apoptose eine komplette Zell**zerstörung**, Gewebe**zerstörung** und damit Tumorzerstörung erreicht werden soll.

English Translation: *Tumors are cooled to -60 °C to -70 °C using cryoprobes, which is expected to result in complete cell* ***destruction****, tissue* ***destruction****, and thus tumor destruction via the development of coagulation necrosis and apoptosis.*

Prompt 1: »Sie haben ein Modell entwickelt, das Koordinationsellipsen in Sätzen erkennt und auflöst. Das Modell gibt fünf verschiedene Versionen des ursprünglichen Satzes zurück, wobei die erste Version die wahrscheinlichste ist. Bitte lesen Sie sich die erste Version des Satzes sorgfältig durch und entscheiden Sie, ob diese Version korrekt ist und den ursprünglichen Satz mit aufgelösten Koordinationsellipsen wiedergibt. Bitte antworten Sie mir nur 'Ja' oder 'Nein' und keiner Erklärung!

Ursprünglicher Satz: 'Die Tumoren werden mittels Kryosonden auf -60 °C bis -70 °C gekühlt, wodurch über die Herausbildung einer Koagulationsnekrose und Apoptose eine komplette Zell-, Gewebe- und damit Tumorzerstörung erreicht werden soll.'

Erste Version: 'Die Tumoren werden mittels Kryosonden auf -60 °C bis -70 °C gekühlt, wodurch über die Herausbildung einer Koagulationsnekrose und Apoptose eine komplette Zellstörung, Gewebestörung und damit Tumorzerstörung erreicht werden soll.'«

Answer: »Nein.«

Prompt 2: »Sie haben entschieden, dass die erste Version des Satzes, die vom Modell als die wahrscheinlichste ausgewählt wurde, nicht korrekt ist und den ursprünglichen Satz mit aufgelösten Koordinationsellipsen nicht vollständig wiedergibt. Das Modell gibt vier weitere Versionen des Satzes zurück, die als die nächst wahrscheinlichsten Versionen ausgewählt wurden. Bitte lesen Sie sich diese vier Versionen sorgfältig durch und wählen Sie die Version aus, die Ihrer Meinung nach am besten den ursprünglichen Satz mit aufgelösten Koordinationsellipsen wiedergibt.

2) 'Die Tumoren werden mittels Kryosonden auf -60 °C bis -70 °C gekühlt, wodurch über die Herausbildung einer Koagulationsnekrose und Apoptose eine komplette Zellzerstörung, Gewebestörung und damit Tumorzerstörung erreicht werden soll.'
3) 'Die Tumoren werden mittels Kryosonden auf -60 °C bis -70 °C gekühlt, wodurch über die Herausbildung einer Koagulationsnekrose und Apoptose eine komplette Zellzerstörung, Gewebezerstörung und damit Tumorzerstörung erreicht werden soll.'
4) 'Die Tumoren werden mittels Kryosonden auf -60 °C bis -70 °C gekühlt, wodurch über die Herausbildung einer Koagulationsnekrose und Apoptose eine komplette Zellversorgung, Gewebestörung und damit Tumorzerstörung erreicht werden soll.'
5) 'Die Tumoren werden mittels Kryosonden auf -60 °C bis -70 °C gekühlt, wodurch über die Herausbildung einer Koagulationsnekrose und Apoptose eine komplette Zellzerstörung, Gewebeerstörung und damit Tumorzerstörung erreicht werden soll.'

Bitte antworten Sie nur mit der richtigen Nummer und ohne Erklärung.«

Answer: »3«

Result: Die Tumoren werden mittels Kryosonden auf -60 °C bis -70 °C gekühlt, wodurch über die Herausbildung einer Koagulationsnekrose und Apoptose eine komplette Zellzerstörung, Gewebezerstörung und damit Tumorzerstörung erreicht werden soll.

English Translation of Prompt Template 1: *»You have developed a model that recognizes and resolves coordination ellipses in sentences. The model provides five different versions of the original sentence, with the first version being the most probable. Please carefully read the first version of the sentence and decide whether this version is correct and represents the original sentence with resolved coordination ellipses. Please respond with either 'Yes' or 'No' without any explanation.*

Original sentence: [...]

First version: [...]«

Answer: *»No.«*

English Translation of Prompt Template 2: *»You have determined that the first version of the sentence, which was selected by the model as the most probable, is not correct and does not fully represent the original sentence with resolved coordination ellipses. The model provides four additional versions of the sentence, which were selected as the next most probable versions. Please carefully read through these four versions and select the version that you believe best represents the original sentence with resolved coordination ellipses.*

2) [...]
3) [...]
4) [...]
5) [...]

Please respond with only the correct number and no explanation.«