Localization of English Affective Narrative Generation to German

Johannes Schäfer, Sabine Weber, and Roman Klinger
Fundamentals of Natural Language Processing, University of Bamberg, Germany
{johannes.schaefer,sabine.weber,roman.klinger}@uni-bamberg.de

Abstract

Prompting strategies for natural language processing with large language models have primarily been developed for English, the most resource-rich language. This raises critical questions about the importance of localization in applications for other languages, i.e. aiming to generate linguistically relevant and culturally appropriate output. In this paper, we investigate the impact of various localization steps within an affective narrative generation task. We experiment with different adaptations, such as the modification or translation of prompts and the usage of localized seed words. We evaluate their influence on the quality of generated texts in German, as compared to English. Our results indicate that a considerable degree of localization is required; however, fully localized prompts do not yield the most natural or localized texts. Additionally, we find that a multilingual model performs well in the simpler aspects of affective narrative generation within the localized context, but it encounters difficulties in the more complex components of the task in German. Our work underscores the complexities involved in adapting large language models for localized emotion analysis with tailored methodologies that account for linguistic variations in narrative generation.

1 Introduction

The advancement of large language models (LLMs) has predominantly emerged from research on English-language data. This has resulted in prompting strategies and methodologies that are largely tailored to this language due to the substantial volume of training data available. This focus raises important questions about the necessity of localization when applying these models to other languages. As LLMs become increasingly multilingual, it becomes crucial to understand whether methods developed for English can be effectively adapted to non-English contexts. Furthermore, it is important to determine whether findings from English-centric research can be generalized to other languages.

Guilt: Ich half meinem Großvater, das Zelt sorgfältig aufzubauen. Er bat mich, den Campingstuhl speziell für ihn aufzustellen. Mein Handy klingelte und lenkte mich von der Aufgabe ab. Ich überhastete das Aufstellen des Stuhls, ohne ihn zu sichern. *Der Campingstuhl kippte um.* (I helped my grandfather set up the tent carefully. He asked me to place the camping chair specifically for him. My phone rang and distracted me from the task. I hastily set up the chair without securing it. *The camping chair tipped over.*)
Surprise: Ich wählte einen ebenen Platz für den Campingplatz aus. Die Zeltstangen wurden doppelt überprüft und fest verankert. Mein Freund und ich planten einen gemütlichen Tag ohne anstrengende Aktivitäten. Wir stellten unsere Getränke und Snacks sorgfältig auf den Campingtisch. *Der Campingstuhl kippte um.* (I chose a flat spot for the campsite. The tent poles were double-checked and firmly anchored. My friend and I planned a cozy day without strenuous activities. We carefully placed our drinks and snacks on the camping table. *The camping chair tipped over.*)

Table 1: German examples with English tranlations of generated stories designed to evoke guilt and surprise.

Our research addresses the localization challenges associated within a specific use case: affective narrative generation. A previous study has been conducted on this task using English language data (Schäfer and Klinger, 2025). This task addresses the ambiguity in emotion analysis on event descriptions by aiming to disambiguate emotions through the generation of contextual information. It comprises two subtasks: first, creating a diverse set of target event descriptions; second, producing contexts that precede these descriptions and influence emotion analysis in a specific way.

To illustrate this task, consider Table 1 with an example ambiguous event description "Der Campingstuhl kippte um." (*The camping chair tipped over.*).[1] When viewed in isolation, the emotion

[1]We provide our annotated dataset and code on https://www.uni-bamberg.de/en/nlproc/resources/german-affective-narrative-generation/.

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 241–256, 2025

evoked in the experiencer is unclear as the situational context is unknown. The first subtask is to generate a varied set of such event descriptions.

The second subtask involves crafting contextual narratives (backstories) that evoke specific emotions in these events. In the example for the case of guilt (see Table 1), the context emphasizes the importance of the situation and especially the personal responsibility for the negative outcome. In the case of surprise, the context highlights the expected nature and carefulness of the previous events, enhancing the perception of abruptness in the event.

This task setup enables us to explore localization in both straightforward text generation, exemplified by producing varied event descriptions, and more complex challenges, such as crafting narratives that strategically influence emotion analysis. This task is especially well-suited for our investigation of localization as it inherently includes a cultural dimension; event descriptions within narratives reflect personal experiences that vary across different cultural and linguistic contexts.

This paper investigates the necessary steps for localizing English-centered methodologies to produce high-quality, contextually appropriate data in German. Our aim is to adapt existing practices in text generation not only to evaluate their effectiveness in a German context but also to understand the unique linguistic features that may influence emotion analysis. Specifically, we focus our investigation on the impact of various localization strategies, as illustrated through the narrative generation process. Our paper is guided by the following key research questions:

1. **What adaptations are necessary to tailor the English event generation approach for the German language, in order to ensure the production of high-quality localized data?** This question aims to identify the complexities involved in adapting methodologies for distinct linguistic contexts, specifically focusing on the necessary modifications to prompts for effective localization.
2. **In the context of generating backstories for events that together form coherent narratives, do the advantages of story planning and narrative revision observed in English also manifest in the German adaptation of the process?** This question seeks to deepen our understanding of narrative story generation techniques and assess the performance of LLMs across different languages.
3. **To what extent do the generated German backstories effectively disambiguate emotion analysis in events, compared to findings from the English-language experiments?** This question explores the role of context in emotion analysis within the German language, aiming to determine whether the trends observed in English are consistent.

2 Related Work

2.1 Localization of LLMs

Recent advancements in text analysis with LLMs continue to highlight the need for effective prompting strategies, which poses particular challenges in diverse linguistic contexts. As noted by Mondshine et al. (2025), the overwhelming dominance of English in LLM research raises questions on the applicability of approaches for other languages and the required adaptations in prompt construction.

The demand for the usage of LLMs in localized applications has led to two main approaches: (1) developing dedicated monolingual models for low-to-medium-resource languages (Seker et al., 2022; Andersland, 2024) and (2) creating multilingual models that utilize data from various languages (Qin et al., 2023; Jiang et al., 2024). Multilingual models can perform well in cross-lingual tasks because of their training on diverse datasets (Raffel et al., 2020; Conneau et al., 2020; Chowdhery et al., 2023) and prompt engineering methodologies that enable effective adaptations (Brown et al., 2020).

To optimize multilingual LLM performance, several prompting techniques have emerged. Huang et al. (2023) introduced XLT, a cross-lingual prompt format for specific languages. Furthermore, Zhao and Schütze (2021) showed that discrete and soft prompting can allow non-English prompts to surpass traditional fine-tuning in cross-lingual contexts. Research on pre-translation techniques has investigated fully translating prompts into English (Chowdhery et al., 2023; Qin et al., 2023) as well as selectively translating specific parts (Liu et al., 2025). Despite their promise, these strategies require further research to establish best practices.

2.2 Emotion Analysis Adaptations

De Bruyne (2023) points out the challenges in applying multilingual approaches due to the culture-specific nature of emotions, a factor critical to our investigation of localization. Fu et al. (2022) demonstrate that training on larger English datasets

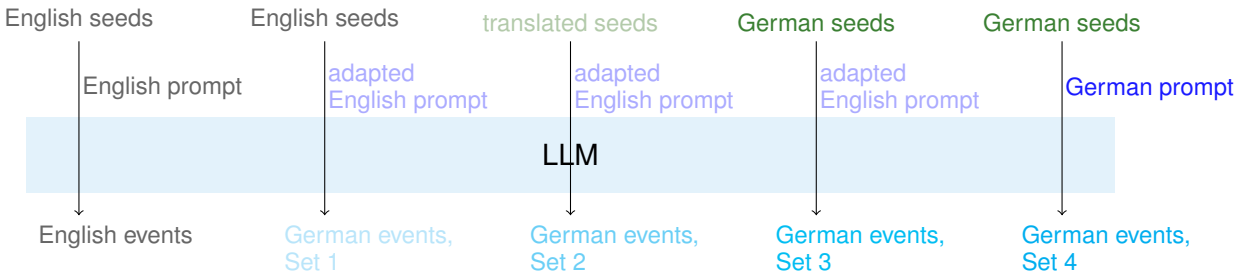

Figure 1: Overview of adaptations of the event generation approach. Items used in the original method by (Schäfer and Klinger, 2025) for generating English data are highlighted in gray text. New items for generating German data are indicated in blue, green, and teal text, with color saturation reflecting the presumed degree of localization.

can benefit both in-language and cross-lingual scenarios, reinforcing the relevance of English-centered methodologies for other languages. In prompting LLMs for emotion analysis, Bareiß et al. (2024) find that using English prompts for non-English texts can enhance classification accuracy.

In the context of narrative generation, understanding emotional dynamics is essential. Mohammad (2012) and Troiano et al. (2023) emphasize the significance of context in emotional categorization. Previous research has begun to explore the role of context in shaping emotional interpretations, with works such as Mostafazadeh et al. (2016) introducing datasets for story completion tasks.

Our research aims to contribute to the understanding of localized prompting strategies within the use-case of affective narrative generation, identifying the critical factors that can clarify emotion analysis and ensure narrative quality.

3 Methods

To explore the impact of localization of LLM-based prompting approaches developed for English, we design a structured approach to evaluate various adaptation strategies for German. Our objective is to assess the effectiveness of localization strategies in the context of affective narrative generation, which consists of two subtasks: first, producing high-quality event descriptions, and second, creating contexts that enhance emotion disambiguation.

We adopt the methodology from Schäfer and Klinger (2025), who experiment on English language data. We therefore start by generating a lexically diverse corpus of event descriptions in German (see Section 3.1), which then serves as the basis for crafting backstories that provide essential context in emotion analysis (see Section 3.2). This approach allows us to examine the effects of specific localizations in prompt[2] construction on lexical diversity, narrative coherence and emotion analysis. In the following, we present modifications to the text generation methods as well as approaches to evaluate the resulting data (see Section 3.3).

3.1 Generation of Events

The English event generation approach by Schäfer and Klinger (2025) uses prompt attributes to aim for diversity and balance of topics. Examples of these attributes (event types and objects) which we also use in our experiments are given in Appendix B. Additionally, ten example events are included within the prompt.

We implement the adaptation of this approach in four distinct ways, each varying in the level of modification and localization applied to the prompts used to guide the LLM's output. For all approaches we translate the ten example event descriptions into German. The different approaches are outlined in Figure 1 and described as follows:

1. **Modification of Prompts:** In this approach, we utilize the original English event types and objects. We modify the prompts by appending "in German" to the instructions. This approach aims to retain the original context while soliciting responses in the target language.
2. **Localized Translation of Categories and Objects:** Building upon the first method, we continue to use the English prompts with the addition of "in German". We translate both the event categories and the corresponding objects into German. This method aims at fostering a more localized response from the LLM by providing German-specific vocabulary.
3. **Creation of Localized Seed Data:** For this adaptation, we retain "in German" in the En-

[2]We provide the original full text prompts used by Schäfer and Klinger (2025) in Appendix A.

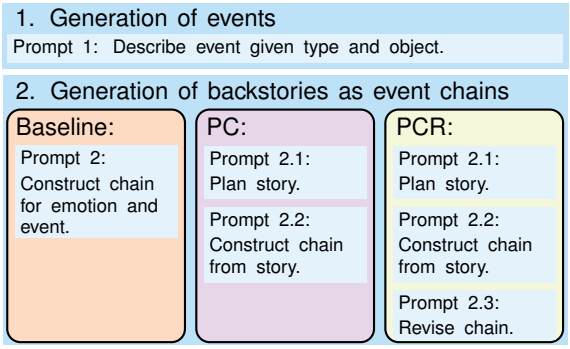

Figure 2: Overview of LLM-based data generation framework presented in Schäfer and Klinger (2025) which we localize for our experiments. The prompts used in each case are shown summarized; full text prompts are given in Appendix A.

glish prompts. Moreover, we employ ChatGPT (OpenAI, 2025) to generate new seed event categories and objects in German.

4. **Fully Localized Generation from Scratch:** In this approach, we also use the newly generated German seed event categories and objects. Additionally, the prompt for generating events is translated into German, in order to aim for outputs that are naturally in line with the German linguistic and cultural context, rather than somewhat translations of English content.

Our hypothesis is that the progression from Method 1 to Method 4 yields increasingly natural German and localized outputs. However, we also hypothesize that as the comparability with the English results may diminish.

3.2 Generation of Backstories

Schäfer and Klinger (2025) present three methods for generating context to disambiguate emotion interpretations for English event descriptions. When faced with an event description where the emotion interpretation may be ambiguous, they generate backstories in order to reduce this ambiguity and steer the interpretation toward a specific emotion.

Figure 2 visualizes the general approach employed in these methods, which are based on varying numbers of prompt interactions with an LLM. The Baseline method utilizes a singular prompt to describe the generation task instruction given an event description and a specific target emotion. The second approach Plan–Construct (PC) builds on this, but adds a story planning step. Furthermore, the third approach Plan–Construct–Revise (PCR) adds another step for revising the generated event chain while reassessing the interpretation. Schäfer and Klinger (2025) have shown that both the story planning and the revision steps are beneficial for generating English data of higher quality. We also apply these three methods to our data in order to find out if this also is the case for German.

We test two adaptation methodologies to generate localized contextual narratives:

1. We minimally modify the three methods by incorporating instructions in the prompts that the output should be "in German".
2. We fully translate the prompts, including the emotion label set, into German to aim for comprehensive localization.

With these two approaches, we evaluate the capabilities of the LLM in performing affective narrative generation in German as compared to English.

3.3 Evaluation

In our experiments, we utilize a multilingual LLM (`meta-llama/Llama-3.3-70B-Instruct`, Meta, 2024) for text generation and a smaller model fine-tuned on German data (`DiscoResearch/Llama3-German-8B`, Disco Research, 2024) for analysis. We chose these models for several key reasons: both are freely available for research and do not rely on API access, which helps eliminate costs and enhances the reproducibility of our experiments. The multilingual LLM is substantial but not overly large, providing a good balance between potential data generation quality and computational cost. Furthermore, the recency and established reputation of both models fosters comparability.

We assess the quality of generated events via human annotation, focusing on naturalness and plausibility (guidelines see Appendix C). For automatic evaluation, we utilize the German LLM to compute transition log likelihood scores across the token sequences of generated texts. We determine the quality of narratives by means of story coherence as shown in Schäfer and Klinger (2025), using a zero-shot shuffle test as proposed by Laban et al. (2021). This test computes token sequence likelihood scores of event chains (we compute these using the German LLM) and compares the values against those of various shuffled permutations of the sequence. Chains that rank highly in this comparison are assigned a high coherence score, whereas those that rank lower receive a low score.

Additionally, we conduct emotion analysis through probabilistic zero-shot classification as described in Schäfer and Klinger (2025), to quantify the emotion content within the generated texts.

Joy: Ich sah den Campingstuhl jeden Tag im Garten stehen. Der Campingstuhl erinnerte mich an den gescheiterten Ausflug mit Freunden. Ich wollte den Stuhl wegwerfen, aber nie die Zeit dafür finden. Ich überlegte, wie ich endlich den alten Stuhl loswerden konnte. *Der Campingstuhl kippte um.*
(I saw the camping chair standing in the garden every day. The camping chair reminded me of the failed trip with friends. I wanted to throw the chair away but never found the time to do so. I contemplated how I could finally get rid of the old chair. *The camping chair tipped over.*)

Boredom: Die Ankunft am Campingplatz erfolgte bei trübem Wetter. Der Zeltplatz wurde täglich auf- und abgebaut. Die gesamte Ausrüstung war bereits mehrmals aufgebaut worden. Keine der vorherigen Campingtage brachte besondere Ereignisse. *Der Campingstuhl kippte um.*
(The arrival at the campsite took place under gloomy weather. The tent site was set up and taken down daily. The entire equipment had already been set up multiple times. None of the previous camping days brought any notable events. *The camping chair tipped over.*)

Table 2: German examples with English tranlations of generated stories evoking joy and boredom.

Id	Set of Events	Length	TTLL
0	English events	17.6	-3.91 $\sigma=0.65$
1	German events, Set 1	14.9	-3.05 $\sigma=0.64$
2	German events, Set 2	15.8	-2.83 $\sigma=0.51$
3	German events, Set 3	14.5	-2.68 $\sigma=0.50$
4	German events, Set 4	23.5	-2.15 $\sigma=0.38$

Table 3: Average text length in number of tokens (Length) and average token transition log likelihood score (TTLL) in event descriptions, averaged in a set.

Ev. Set:	0		1		2		3		4	
Annot.	N	P	N	P	N	P	N	P	N	P
#1	4.16	4.96	4.33	4.79	4.44	4.64	4.68	4.72	4.56	4.76
#2	4.00	4.96	3.75	5.00	4.44	4.92	4.12	4.92	3.80	4.88
#3	4.84	4.92	4.38	4.96	4.68	4.84	4.76	4.88	4.60	4.88
Mean	4.33	4.95	4.15	4.92	4.52	4.80	4.52	4.84	4.32	4.84

Table 4: Average annotated 1–5 Likert scale values of naturalness (N) and plausibility (P) in events sets (25 samples per set) as rated by annotators #1, #2 and #3.

4 Results

This section presents the findings aligned with our research questions regarding the impact of localization in our use case of affective narrative generation. We display selected examples of generated narratives in Table 1 and Table 2. We begin with Section 4.1, addressing our first research question about the effectiveness of various localized prompts in event generation. Next, Section 4.2 targets our second question by evaluating the backstories produced through different approaches. Finally, in Section 4.3, we analyze the emotions evoked by the generated events with and without context, corresponding to our third research question.

4.1 Localization of Generated Events

Our goal is to generate a corpus of event descriptions as basis for emotion analysis experiments. Each of the methods for adapting the approach from English to German (as described in Section 3.1) generates a set of 1000 events. We now evaluate the results from the application of these methods in order to answer our first research question: **What adaptations are necessary to tailor the English event generation approach for the German language, in order to ensure the production of high-quality localized data?** We automatically evaluate the entire dataset and manually assess nuances in perceived quality, addressing potential limitations of the automatic approach.

Automatic Evaluation. We utilize the German LLM to calculate the likelihood of the token sequence in each event description produced. The overall results for each set of events are summarized in Table 3. Our findings show a clear correlation between the degree of adaptation or translation of the methods employed and the resulting transition probabilities of the token sequences. As the methods become more tailored to the German language, the transition probabilities increase. Additionally, the noticeable decrease in standard deviation supports this observation, showing that the adaptations increasingly produce more consistently high-quality German language data.

Our results verify the language model's preference for German, as it shows a substantial inclination toward all German event descriptions compared to the English texts (en events). We also observe a substantial enhancement in the transition probability for the last set of events, which was generated using prompts exclusively in German.

Human Annotation. We now analyze a subset of the generated data to assess whether human evaluators align with our automatic evaluations regarding the quality of the language produced. We ask three annotators to rate the naturalness (linguistic quality) and plausibility (semantics) of the texts. Details of the procedure as well as inter-annotator agreement values are shown in Appendix C.

Table 4 summarizes the annotation results of the

naturalness and plausibility assessment. The events generated in Set 2 and Set 3 receive the highest naturalness ratings on average. Conversely, the events in Set 1, which involves only minimal localization to the German context, fall substantially short, underscoring the need for a substantial adaptation of the original approach. Interestingly, the texts generated using prompts composed entirely in German (Set 4) do not achieve the highest naturalness ratings. This shows that complete localization does not correlate with perceived naturalness in human evaluation. For the semantic plausibility assessment, we observe minimal differences across methods. All approaches receive high scores, indicating their ability to generate plausible events.

These results warrant a summary to address our first research question. Our evaluation shows that to effectively tailor the English event generation approach for the German language, specific adaptations are essential. Firstly, substantial localization of seed words is necessary to enhance the relevance and contextual appropriateness of the generated events. This is evidenced by the marked performance differences between Set 1 and the more localized Sets 2, 3, and 4. Moreover, while our findings show that using prompts in German (as observed in Set 4) does not lead to higher naturalness ratings, it highlights a nuanced relationship between localization and perceived text quality. Therefore, we recommend that future work should explore both approaches, while also conducting thorough human evaluations of the results to identify the best strategy for specific tasks.

4.2 Narrative Content of Backstories

We now shift our focus to the generated contextual narratives based on the events produced using the methods outlined in Section 3.2. Our goal is to address our second research question: **In the context of generating backstories for events that together form coherent narratives, do the advantages of story planning and narrative revision observed in English also manifest in the German adaptation of the process?** We analyze the lexical content of the backstories organized by the respective emotion categories, examine the differences among the generation methods, and explore the relationships between the two prompting setups.

Setting. As basis we use the two sets of German events which scored highest in both automatic and human evaluations from the previous section (Set 3 and 4). Each set consists of 1,000 event descriptions. For each event description, 13 backstories comprising four preceding events are generated, with each backstory crafted to evoke one of 13 emotion categories. The experimental design for backstory generation distinguishes between the two sets based on their initial generation prompts. For Set 3, generated using English prompts, we generate backstories with English prompts and the English emotion set. Conversely, Set 4 employs German prompts, with backstories generated with German prompts and the German emotion set. This setup allows us again to test whether prompting LLMs in English may yield superior results compared to those prompted in German, while also assessing the impact of localization.

With this approach, we generate backstories corresponding to each of the 1,000 German events in both sets using three different prompting methods as discussed in Section 3.2.

Content of Event Chains. To gain insights into the narratives of these backstories, particularly at a lexical level, we now focus on identifying the most prominent words – specifically nouns – by determining their term frequency-inverse document frequency (TF-IDF) values. Results of the comprehensive analysis is provided in Appendix D. We summarize selected findings in the following.

Notably, the results show clear distinctions between the subsets corresponding to their respective emotions, as expected. For instance, event chains associated with disgust are predominantly characterized by themes related to scandals, corruption, and waste, while those concerning sadness reflect more familial themes such as grandpa, grandma, and grave. In the context of fear, terms like unrest, AI and caution are frequently observed.

When comparing the three methods of backstory generation, an initial trend emerges: backstories generated with the PC and PCR approaches tend to exhibit more complex associations with the respective emotions compared to those generated by the Baseline approach. This becomes particularly evident when we examine the absolute TF-IDF scores, which illustrate that the PC and PCR approaches create more pronounced lexical distinctions among emotions by utilizing more specific vocabulary.

Comparing the data from both prompt setups, we observe that the localized German prompts produce longer compound nouns, while data generated using English prompts lean towards shorter words.

		Mean Coherence of Chains		
		Baseline	PC	PCR
Emotion-Specific Subsets	Anger	.74	.76 $\Delta+.02$	.77 $\Delta+.03$
	Boredom	.75	.75 $\Delta\pm.00$	.77 $\Delta+.02$
	Disgust	.72	.78 $\Delta+.06$	.79 $\Delta+.07$
	Fear	.71	.73 $\Delta+.02$	.77 $\Delta+.06$
	Guilt	.76	.79 $\Delta+.03$	.81 $\Delta+.05$
	Joy	.77	.77 $\Delta\pm.00$	.79 $\Delta+.02$
	Pride	.78	.80 $\Delta+.02$	.81 $\Delta+.03$
	Relief	.78	.76 $\Delta-.02$	.81 $\Delta+.03$
	Sadness	.74	.79 $\Delta+.05$	.82 $\Delta+.08$
	Shame	.75	.80 $\Delta+.05$	.78 $\Delta+.03$
	Surprise	.74	.76 $\Delta+.02$	.80 $\Delta+.06$
	Trust	.77	.78 $\Delta+.01$	.78 $\Delta+.01$
	No-Emotion	.71	.70 $\Delta-.01$	.75 $\Delta+.04$
Overall Dataset		.75 $\sigma = 0.26$	.77 $\Delta+.02$ $\sigma = 0.26$	.79 $\Delta+.04$ $\sigma = 0.25$

Table 5: Mean coherence scores for event chains in the datasets generated with different methods based on event Set 3. Δ values show the difference of the respective score in comparison to the Baseline.

		Mean Coherence of Chains		
		Baseline	PC	PCR
Emotion-Specific Subsets	Wut	.80	.76 $\Delta-.04$	.80 $\Delta\pm.00$
	Langeweile	.77	.75 $\Delta-.02$	.78 $\Delta+.01$
	Ekel	.78	.77 $\Delta-.01$	.80 $\Delta+.02$
	Angst	.76	.72 $\Delta-.04$	.77 $\Delta+.01$
	Schuld	.79	.79 $\Delta\pm.00$	.81 $\Delta+.02$
	Freude	.81	.78 $\Delta-.03$	.81 $\Delta\pm.00$
	Stolz	.82	.77 $\Delta-.05$	.80 $\Delta-.02$
	Erleichterung	.81	.77 $\Delta-.04$	.81 $\Delta\pm.00$
	Traurigkeit	.81	.77 $\Delta-.04$	.81 $\Delta\pm.00$
	Scham	.81	.78 $\Delta-.03$	.82 $\Delta+.01$
	Überraschung	.80	.78 $\Delta-.02$	.81 $\Delta+.01$
	Vertrauen	.79	.76 $\Delta-.03$	.79 $\Delta\pm.00$
	Keine_Emotion	.79	.76 $\Delta-.03$	.78 $\Delta-.01$
Overall Dataset		.79 $\sigma = 0.24$	.77 $\Delta-.02$ $\sigma = 0.25$	.80 $\Delta+.01$ $\sigma = 0.24$

Table 6: Mean coherence scores for event chains in the datasets generated with different methods based on event Set 4. Δ values show the difference of the respective score in comparison to the Baseline.

Coherence Analysis. Further, we perform a coherence analysis as described in Section 3.3. Here, we evaluate the coherence scores two times in three sets of generated event chains: one generated with English prompts based on event Set 3 (displayed in Table 5) and the other generated with German prompts based on event Set 4 (shown in Table 6). This analysis is further refined by differentiating between the three approaches and the respective emotion categories evoked in the target events.

In Table 5, the overall coherence of chains generated using the baseline approach is observed to be approximately .75. The introduction of story planning, included in the PC approach, results in an increase to a coherence score of .77. Further revisions, incorporated in the PCR approach, yield an even higher coherence of .79.

When examining the coherence of data generated with German prompts (see Table 6), we find that baseline chains achieve a relatively high coherence score of .79. However, this time the integration of story planning correlates with a decrease in coherence to .77. Only with additional revisions do we achieve the highest overall average coherence score of .80.

While there are no extreme outliers across specific emotions, we observe a trend indicating that the categories of boredom, fear, and no-emotion tend to exhibit the least coherent event chains.

Taking these results into account, we now address our second research question regarding the effect of story planning and narrative revision in the German adaptation of the event generation process. Our findings show that story planning contributes to the generation of lexically richer and better localized backstories. However, it does not consistently lead to higher coherence in the narratives, as observed in certain cases where coherence scores decrease after incorporating story planning. Nevertheless, when paired with narrative revision, we observe an improvement in coherence, achieving the highest levels when both techniques are applied. Comparatively, the improvements seen in German narratives are less pronounced than those noted by Schäfer and Klinger (2025) in English. This shows that while benefits from planning and revision exist, their effects are not as pronounced in German as compared to English using the multilingual LLM.

4.3 Emotion Analysis

With the localized sets of events and generated narratives, we now perform a systematic and comprehensive analysis of the effect of context in emotion analysis. Here we address our third research question: **To what extent do the generated German backstories effectively disambiguate emotion analysis in events, compared to findings from the English-language experiments?**

Setting. We perform automatic emotion analysis as described in Section 3.3: We first examine the prediction probabilities on the events in isolation, and then contrast these with the probabilities when contextual narratives from the backstories are in-

Event Set Id	Anger	Boredom	Disgust	Fear	Guilt	Joy	Pride	Relief	Sadness	Shame	Surprise	Trust	No-Emotion
0	.02	.01	.00	.01	.00	.33	.01	.08	.01	.00	.14	.02	.37
3	.01	.00	.01	.02	.00	.37	.07	.15	.09	.00	.06	.03	.20
4	.01	.00	.00	.02	.00	.51	.15	.10	.07	.00	.04	.06	.03

Table 7: Results $p(e|E)$ of probabilistic zero-shot emotion analysis for each emotion (e) on average given the events E from different event sets.

cluded. With this comparison we determine if the added contexts disambiguate emotion analysis.

Emotion Analysis of Events. In Table 7, we compare the results for automatic emotion analysis on events in two German event sets (Set 3 and 4) to those for the English event set introduced by Schäfer and Klinger (2025) (Set 0). Overall, these results are in line with an expected prior distribution of emotions in event descriptions. However, several notable differences emerge from a comparison between the sets. Events from the German sets evoke stronger emotions; the probability of detecting the no-emotion category is considerably lower for the German sets (.20 and .03) compared to the English data (.37). This increased emotionality affects only certain emotion categories that are more prevalent in the German data, particularly joy, pride, and sadness. Conversely, the German events encompass the emotion of surprise less frequently.

Emotion Analysis in Contexts. Next, we conduct an automatic emotion analysis on the event chains to ascertain whether the addition of generated backstories successfully amplifies the evocation of specific emotions in the target events, thus contributing to the disambiguation of emotion analysis. We again analyze two times three sets of event chains: one evaluated with emotion analysis using English prompts applied to event Set 3 (see Table 8) and another assessed using German prompts and relevant emotion sets on event Set 4 (see Table 9).

The results illustrated in Table 8 show that the emotions joy, pride, relief, and sadness are particularly well-fostered by their corresponding backstories, especially when compared against the initial emotions of the events, as presented in Table 7. The added context has a substantial influence on these emotions. However, when comparing the various approaches – Baseline, PC, and PCR – no systematic differences emerge overall in their effectiveness on emotion evocation. Comparatively, examining the data generated with German prompts (see Table 9) reveals that the evoked emotions are less influenced by the added contexts. Although the trends across categories are consistent, the overall values are notably lower.

Reflecting on the overall results, we can now answer our third research question concerning the effectiveness of the generated German backstories in disambiguating emotion analysis in events, particularly in comparison to the findings from the English-language experiments by Schäfer and Klinger (2025). Our results reveal that while certain emotion categories – including joy, pride, and

		$p(e = e_p\|C)$		
		Baseline	PC	PCR
Emotion-Specific Subsets	Anger	.10	.10 $\Delta\pm.00$	.13 $\Delta+.03$
	Boredom	.02	.07 $\Delta+.05$	.08 $\Delta+.06$
	Disgust	.09	.11 $\Delta+.02$	.12 $\Delta+.03$
	Fear	.17	.16 $\Delta-.01$	.15 $\Delta-.02$
	Guilt	.18	.16 $\Delta-.02$	.16 $\Delta-.02$
	Joy	.34	.31 $\Delta-.03$	.29 $\Delta-.05$
	Pride	.23	.23 $\Delta\pm.00$	.29 $\Delta+.06$
	Relief	.45	.45 $\Delta\pm.00$	.42 $\Delta-.03$
	Sadness	.39	.39 $\Delta\pm.00$	.34 $\Delta-.05$
	Shame	.10	.12 $\Delta+.02$	.12 $\Delta+.02$
	Surprise	.13	.11 $\Delta-.02$	.11 $\Delta-.02$
	Trust	.07	.06 $\Delta-.01$	.06 $\Delta-.01$
	No-Emotion	.01	.02 $\Delta+.01$	.02 $\Delta+.01$
Overall Dataset		.18 $\sigma = .37$	.18 $\Delta\pm.00$ $\sigma = .37$	.18 $\Delta\pm.00$ $\sigma = .37$

Table 8: Prediction emotion probability evoked in the last event from event Set 3 ($p(e)$) given entire event chains (C) with the respective prompted emotion (e_p).

		$p(e = e_p\|C)$		
		Baseline	PC	PCR
Emotion-Specific Subsets	Wut	.02	.04 $\Delta+.02$	.04 $\Delta+.02$
	Langeweile	.02	.01 $\Delta-.01$	.03 $\Delta+.01$
	Ekel	.02	.06 $\Delta+.04$	.08 $\Delta+.06$
	Angst	.05	.08 $\Delta+.03$	.10 $\Delta+.05$
	Schuld	.02	.06 $\Delta+.04$	.08 $\Delta+.06$
	Freude	.39	.42 $\Delta+.03$	.46 $\Delta+.07$
	Stolz	.17	.24 $\Delta+.07$	.31 $\Delta+.14$
	Erleichterung	.19	.25 $\Delta+.06$	.26 $\Delta+.07$
	Traurigkeit	.19	.25 $\Delta+.06$	.33 $\Delta+.14$
	Scham	.02	.04 $\Delta+.02$	.05 $\Delta+.03$
	Überraschung	.08	.06 $\Delta-.02$	.07 $\Delta-.01$
	Vertrauen	.05	.08 $\Delta+.03$	.09 $\Delta+.04$
	keine_Emotion	.01	.00 $\Delta-.01$	.00 $\Delta-.01$
Overall Dataset		.09 $\sigma = .26$	.12 $\Delta+.03$ $\sigma = .30$	.14 $\Delta+.05$ $\sigma = .32$

Table 9: Prediction emotion probability evoked in the last event from event Set 4 ($p(e)$) given entire event chains (C) with the respective prompted emotion (e_p).

sadness – show improved disambiguation when contextualized with the narratives, for the other emotions this effect is less pronounced in German than in English. Specifically, the German backstories do foster emotion evocation only to some degree; however, coherence is possibly being prioritized over precise portrayal of emotions in the narrative generation process. Moreover, the expected benefits of the multi-step data generation techniques do not translate as clearly in the results for emotion analysis in German. This highlights that the task is more effectively handled in English.

5 Conclusion

With this paper, we provide an investigation of the localization of English LLM-prompting methodologies to German, highlighting the role of contextual disambiguation in narratives. Our findings provide key insights into the complexities of adapting emotion analysis techniques across languages.

Firstly, we now understand the adaptations necessary for customizing event generation techniques to German. Our results show that substantial localization, particularly of seed words, is essential for improving the relevance of generated events. It is necessary to balance approaches that account for both localization and text quality.

Secondly, our evaluation of narrative generation shows that while story planning results in lexically richer backstories, it does not consistently improve coherence in generating German data. We find that while the multilingual LLM performs adequately in simpler generation tasks, it struggles with the more complex components of narrative coherence and the disambiguation of emotion analysis in German compared to English. Combining story planning with narrative revision does lead to notable benefits; however, these improvements are less pronounced in German. These results show that the multilingual LLM may not be able to perform complex generation tasks in languages other than English as effectively. Therefore, specific language fine-tuning is advisable to enhance performance in these tasks and ensure more accurate and localized outputs.

Lastly, our exploration of the generated backstories' effectiveness in disambiguating emotions in German event descriptions is only effective for certain emotions. This highlights the need for further research into the adaptation of prevalent English data generation approaches, particularly in managing complex narrative requirements.

Our findings prompt several essential follow-up questions. Firstly, how can the emotion disambiguation technique be further adapted for other languages, particularly those with significant grammatical and lexical differences from English? Additionally, what specific linguistic nuances within German and other languages might affect the effectiveness of the localization? Furthermore, there remains a need to explore the scalability of our approach across various narrative contexts and cultures, which could unveil new strategies for enhancing emotion analysis in contextualized settings. Addressing these questions would refine our methodologies and contribute to the ongoing discourse on the applicability of large language models in diverse linguistic environments, particularly in languages underrepresented in LLM pre-training data.

In summary, our research emphasizes the challenges of adapting established approaches with prompting LLMs to generate English data to less represented languages. In particular, research of localized phenomena, such as the contextual influence in emotion analysis, tends to still be challenging with multilingual LLMs. Future work should continue to refine methodologies and evaluation frameworks to better accommodate the unique characteristics of the German language.

Limitations

This study has several limitations that may impact the generalizability of our findings. Firstly, we employed only one LLM for text generation, which restricts our ability to assess the effectiveness of localization strategies across different models. Future research should explore multiple LLMs to validate our results. We provide our code and data to facilitate reproducibility. Additionally, our focus on German limits the applicability of our findings to languages with distinct grammatical and cultural characteristics. The localization strategies identified may not transfer directly to such languages. Lastly, our emotional analysis was based on a limited set of emotions, which may overlook critical nuances in emotional expression. Expanding this scope could yield a more comprehensive understanding of localization in emotion analysis.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) in the project KL2869/1–2 (CEAT, project number 380093645).

References

Michael Andersland. 2024. Amharic llama and llava: Multimodal llms for low resource languages. *Preprint*, arXiv:2403.06354.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for nli-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1318–1326, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466, Toronto, Canada. Association for Computational Linguistics.

Disco Research. 2024. Model llama3-german-8b-v0.1. Large language model.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot prompt: Multilingual multitask prompt training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9919–9935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. Can transformer models measure coherence in text: Re-thinking the shuffle test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models. *Preprint*, arXiv:2403.10258.

Meta. 2024. Llama (model llama-3.3-70b-instruct). Large language model.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond English: The impact of prompt translation strategies across languages and tasks in multilingual LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1331–1354, Albuquerque, New Mexico. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

OpenAI. 2025. Chatgpt (model gpt4o-mini). Large language model.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is

ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. 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. *Journal of Machine Learning Research*, 21(140):1–67.

Johannes Schäfer and Roman Klinger. 2025. Shaping event backstories to estimate potential emotion contexts. *Preprint*, arXiv:2508.09954.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. AlephBERT: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56, Dublin, Ireland. Association for Computational Linguistics.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Full Text Prompts

Table 10 shows the full text of the prompts used by Schäfer and Klinger (2025) to generate data according to the different methods. The interaction with the model comprises three message types[3]: a "system" message that establishes the context for the interaction and includes general guidelines; a "user" message that encapsulates the specific inputs, requirements, and instructions for the task; and an "assistant" message that represents the model's response based on the provided context.

Prompt 1 is used to generate the concluding event and aims to ensure a balanced topic variance through the incorporation of specific attributes. To ensure that the generated event descriptions are concise, Schäfer and Klinger (2025) implement few-shot prompting by including ten examples of event descriptions in the prompt.

[3] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/

For each event generated in the first step, Schäfer and Klinger (2025) aim to create a corresponding backstory comprised of four preceding events, tailored to influence the specific emotion responses of the last event. In the baseline approach, this process is implemented using a single prompt (**Prompt 2** as shown in Table 10). For the other methods, the process is further broken down into sub-steps: Story planning using **Prompt 2.1:** For each final event, the LLM is prompted to first generate a story plan that outlines plausible explanations for the emotion responses tied to the concluding event. This preemptive planning is intended to increase coherence and relevance in the backstories. Based on the crafted story plan, the LLM should generate the sequential backstory comprising four events. This ensures that the events are narratively connected and describe the context leading up to the pivotal final event. Event chain generation using **Prompt 2.2:** After generating the event chain, Schäfer and Klinger (2025) suggest to conduct a summarization step to ensure uniformity and clarity. Additionally, the third method by Schäfer and Klinger (2025) uses **Prompt 2.3** to revise the generated event chain in an attempt to improve the adherence to the defined requirements.

Table 11 shows the English prompt messages used for automatic emotion analysis.

B Event Types and Objects

Table 14 shows the list of ten English event types along with ten corresponding examples of related objects. Table 15 shows the list of ten German event types along with ten corresponding examples of related objects. These values serve as attributes in the prompt which initially generates diverse events that we subsequently develop backstories for.

C Details of the Human Annotation

In this section we provide the annotation guidelines used to rate the quality of event descriptions in Appendix C.1 as well as the inter-annotator agreement of the annotated data in Appendix C.2.

C.1 Annotation Guidelines

Given is a list with descriptions of events.

The data is separated by language (English/German). It should be assumed that each text was written by native speakers.

#	MT	Prompt Text
Prompt 1	system	You are a person describing an event which you have experienced. 10 examples of such event descriptions are as follows: 0: The phone rang. 1: A cat meowed. 2: The car engine sputtered to a stop. 3: A child laughed in the park. 4: A bird fluttered past the window. 5: The waves crashed against the shore. 6: A train whistled as it approached. 7: The fireworks lit up the sky. 8: A bicycle rode by. 9: A crowd cheered at the concert.
	user	The event you experienced is of type: {ds_event_type}. In a longer text you are describing several things that happened at that event. Something happened at that event with the following object(s): {ds_event_object}. In your response, only provide a very short sentence describing what happened to/with the object(s).
Prompt 2	system	It is often clear from the text that describes an event which specific emotion it evokes in a person that experienced it. However, additional information about the situation can change our understanding of how a person might interpret the event. You are an expert at creating a scenario that explains why a specific event may cause a possibly unusual emotion in you. In addition, you can concisely make this scenario apparent for the reader by formulating a description of 4 events that took place immediately before the event.
	user	You experienced something happening which is described by the following event description: 5. "{event}". This event somehow made you clearly feel the emotion: "{emotion}". Provide a text describing four events that took place immediately before event 5 by giving a list of descriptions of these events (1.-4.). The events 1.-4. clearly influence your personal emotional interpretation of the event that happened after (5.). The emotion "{emotion}" is only triggered by what specifically happened in event 5. The events 1.-4. evoked other emotions, such as: {", ".join(random.sample([emo for emo in EMOTION_SET if emo != emotion], len(EMOTION_SET)-1))}. In your response, for each of the 4 event descriptions: Only give a summary text consisting of the main clause in a very short sentence. Each description should only describe a singular event. Indicate each event description in a separate line.
Prompt 2.1	system	It is often clear from the text that describes an event which specific emotion it evokes in a person that experienced it. However, additional information about the situation can change our understanding of how a person might interpret the event. You are an expert at creating a scenario that explains why a specific event may cause a possibly unusual emotion in you. In addition, you can concisely make this scenario apparent for the reader by formulating a description of 4 events that took place immediately before the event.
	user	You experienced something happening which is described by the following event description: 5. "event". This event somehow made you clearly feel the emotion: "emotion". First, give a brief explanation of a scenario in which it can be deduced from the description of event 5. that you felt emotion. Second, phrase this explanation as events that took place immediately before event 5 by giving a list of descriptions of these events (1.-4.). The events 1.-4. clearly influence your personal emotional interpretation of the event that happened after (5.). The emotion "emotion" is only triggered by what specifically happened in event 5. The events 1.-4. evoked other emotions, such as: {", ".join(random.sample([emo for emo in EMOTION_SET if emo $\neq$ emotion], len(EMOTION_SET)-1))}.
Pr. 2.2	user	Extract the sequence of 4 descriptions of events that happened from the following text: ### {explanation} ### The event 5: "{event}" happened after the 4 events. In your response, for each of the 4 event descriptions: Only give a summary text consisting of the main clause in a very short sentence. Each description should only describe a singular event. Indicate each event description in a separate line.
Prompt 2.3	system	You are an expert at adapting a narrative to convey specific emotional interpretations. You will receive a text that outlines a sequence of events as experienced by an individual. Additionally, there will be an explanation of how a particular emotion is triggered in this individual based on the final event.
	user	Explanation: {story_plan} Event sequence: {chain} First, provide a brief evaluation on how the first four events (1.-4.) of the sequence could be adjusted to form a coherent narrative which better aligns with the conclusion given in the explanation. The text of the last event (5.) should remain as is. Second, provide a revised event sequence that incorporates these adjustments while keeping the sentence length for each event description similar. Each event description should consist only of a main clause in a very short sentence. Do not explicitly mention the emotions felt.

Table 10: List of Enlish prompts used by (Schäfer and Klinger, 2025) for event chain generation. The first column shows the numbers of the specific prompts as introduced in Figure 2. 'MT' refers to the prompt message type as specified in the input for the instruction-tuned LLM.

For each event, the following should be annotated:

- **Naturalness (linguistic level):** Evaluated on how authentic and idiomatic the sentence sounds. Criteria are:
  - Language flow: Is the sentence fluent to read?
  - Grammar and syntax: Is the sentence grammatically correct?
  - Word choice: Is the word choice appropriate for the event and is the stylistic level correctly chosen?
- **Plausibility (semantic level):** Evaluated on how believable and realistic the event is. Criteria are:
  - Personal perspective: Would someone who experienced the event describe it this way?
  - Reality reference: Is the event realistic?
  - Contextualization: Does the event fit within the cultural context?

Annotation labels are on a scale from 1-5:

1. Not natural/plausible at all,
2. A little natural/plausible,
3. Moderately natural/plausible,
4. Quite natural/plausible,
5. Totally natural/plausible.

MT	Prompt Message Text
System	You are an expert in emotion analysis on event descriptions.
User	A person describes their experience as follows: {text_instance} What emotion was evoked in the person at the end? As your response, provide only one label from the emotion set: anger, disgust, fear, guilt, joy, sadness, shame, pride, boredom, surprise, trust, relief, no-emotion.
Assistant	{emotion}

Table 11: Prompt messages for zero-shot emotion analysis on events and event chains as introduced by Schäfer and Klinger (2025). 'MT' refers to the respective prompt message type as specified in the input for the instruction-tuned LLM.

	#1	#2	#3
#1	0.00	0.75	0.54
#2	0.75	0.00	0.81
#3	0.54	0.81	0.00

Table 12: Mean Δ between naturalness ratings of pairs of annotators.

C.2 Inter-Annotator Agreement

The annotated data comprises 125 event descriptions (25 texts from each of the five sets) which were annotated by three annotators. To assess inter-annotator agreement, we calculate weighted Cohen's Kappa between pairs of annotators as well as Krippendorff's Alpha overall. This results in the following values for the annotation:

- Naturalness (Weighted Cohen's Kappa)
  - #1 – #2: 0.45
  - #1 – #3: 0.34
  - #2 – #3: 0.34
  - Krippendorff's Alpha: 0.31
- Plausibility (Weighted Cohen's Kappa)
  - #1 – #2: 0.09
  - #1 – #3: 0.26
  - #2 – #3: 0.17
  - Krippendorff's Alpha: 0.17

	#1	#2	#3
#1	0.00	0.26	0.25
#2	0.26	0.00	0.14
#3	0.25	0.14	0.00

Table 13: Mean Δ between plausibility ratings of pairs of annotators.

Note that, while these measures for agreement show quite low values here, we should take into account that the label distribution in this dataset is extremely skewed. Most instances were annotated with high naturalness (labels 3-5) and very high plausibility (4-5). In this case, these measures include a very high penalty for chance agreement.

To assess the quality of the annotation with another measure, we include the mean Δ between the annotated scores of two annotators. The results for this assessment are shown in Table 12 for the annotation of naturalness and in Table 13 for the annotation of plausibility. This evaluation shows that, especially for the annotation of plausibility of event descriptions, there is very little variance between the ratings of the annotators.

D TF-IDF-Based Unigram Analysis of Generated Event Chains

This section provides results from our evaluation of the generated event chains by displaying the top words in different subsets of our data sorted by term frequency-inverse document frequency (TF-IDF) values. Table 16 shows the top 10 nouns in the event chains generated using event Set 3, separately for the different English emotion categories used. Table 17 shows the top 10 nouns in the event chains generated using event Set 4, separately for the different German emotion categories used. The separation according to the columns of the tables correspond to the different generation methods Baseline, PC and PCR. The text background color saturation is based on the respective TF-IDF values normalized in the range of the values of all words of each table.

Event Type	Event Type Objects
Social Gathering	chairs, tables, food platters, drinks, napkins, decorations, music speakers, games, invitation cards, host
Educational Activity	textbooks, notebooks, pencils, whiteboard, markers, projector, handouts, calculator, overhead projector, globe
Recreational and Nature Activity	hiking boots, backpacks, water bottles, first-aid kit, campfire supplies, nature guide, binoculars, tent, sleeping bags, camping chairs
Cultural and Community Event	stage, performers, sound system, projector, festival tickets, food stalls, craft booths, cultural displays, artworks, costumes
Professional Development	business cards, presentation slides, notebooks, pens, projector, handouts, networking tools, feedback forms, laptops, name badges
Celebration	cake, candles, party hats, balloons, confetti, party favors, streamers, drinks, gift bags, music playlist
Artistic Performance	stage, costumes, sets, props, lights, sound equipment, musical instruments, audience seats, backdrops, tickets
Competition	trophies, medals, referee kit, scoreboard, team jerseys, game equipment, whistle, competition schedule, event tickets, player registration
Family and Relationships	family photo albums, toys, family tree chart, gift cards, family recipe book, family calendars, cameraman, outdoor equipment, gifts, personalized items
Transportation and Travel Event	maps, itineraries, backpacks, suitcases, boarding passes, tickets, travel guides, snacks, passports, travel pillows

Table 14: List of English event types and objects used in prompt templates to create diverse event descriptions.

Event Type	Event Type Objects
Freizeit-/Naturaktivität	Fahrrad, Wanderschuhe, Zelt, Campingstuhl, Kamera, Angelschnur, Rucksack, Boot, Kajak, Angelrute
Wettbewerb	Medaille, Trophäe, Zertifikat, Pokale, Urkunde, Rangliste, Wettbewerbsbedingungen, Prüfung, Wettbewerbsbeitrag, Teilnahmeformular
Berufliche Erfahrung	Bewerbungsgespräch, Lebenslauf, Arbeitsvertrag, Präsentation, Teammeeting, Fortbildung, Mentoring, Networking-Event, Projektabschluss, Feedbackgespräch
Familie/Beziehung	Eltern, Geschwister, Großeltern, Onkel, Tante, Cousin, Cousine, Ehepartner, Freunde, Kinder
Reiseerlebnis	Reisepass, Koffer, Flughafenticket, Hotelbuchung, Rucksack, Kamera, Wanderkarte, Führerschein, Stadtführer, Reiseapotheke
Lebensverändernde Entscheidung	Wohnungskauf, Hauskauf, Führerschein, Neues Auto, Reisepass, Kreditvertrag, Ehevertrag, Adoptionspapiere, Jobangebot, Rentenbescheid
Kulturelles Erlebnis	Museum, Kunstgalerie, Theaterstück, Konzerthalle, Kulturfestival, Filmfestival, Oper, Buchmesse, Kunstwerk, Tanzaufführung
Herausforderung/ Überwindung	Krankheit, Schwierige Entscheidung, Finanzielle Probleme, Trennung, Mobbing, Arbeitslosigkeit, Umzug, Studienabbruch, Naturkatastrophe, Unfall
Freundschaft	Freund, Freundin, Geschenk, Gemeinsames Essen, Kaffeetasse, Kino, Picknick, Reise, Geburtstag, Chat
Gesundheit/ Wohlbefinden	Vitamin D, Yoga, Meditation, Ernährung, Fitnessstudio, Laufen, Schwimmen, Gesundheitscheck, Stressbewältigung, Seele

Table 15: List of German event types and objects used in prompt templates to create diverse event descriptions.

Subset	Baseline	PC	PCR
Anger	Absprache, Versicherung, Trauerarbeit, Erzfeind, Lieblingsparkplatz, Observatoriums, Teilnehmers, Spielern, Stellenabbau, Respekt	Absprachen, Assistent, Agentur, Gebühr, Konto, Alex, Fotograf, Fokuskriterien, Initiativen, Täuschung	Assistent, Antiquitäten, Ärzten, Begründung, Gebühr, Recruiter, Alex, Lüge, Wasservorrat, Jäger
Boredom	Beschäftigung, Kaffeemaschine, Radfahrer, Betriebsfeier, Lautstärke, Meeresboden, Kajaklehrer, DartPartner, Fischsandwich, Zeltbau	Preisträger, Wiederholung, Fusion, Gewohnheit, Rednerin, Formulare, Ritual, Festivalgelände, Putzen, Organisatorin	Preisträger, Fusion, Schema, Rednerin, Kreditberater, Saals, Jahrhundert, Besuchs, Erzählabende, Unterbrechung
Disgust	Geschäftsmann, Werbeplakaten, Wirklichkeit, Betrüger, Skandal, Substanz, Korruption, Chemikalien, Umweltverschmutzer, Gärtnerei	Praktiken, Skandale, Korruption, Verschmutzung, Substanz, Abfälle, Bestechungsgelder, Umfragen, Beweise, Handschuhe	Korruption, Chemikalien, Verschmutzung, Substanz, Lieblingskandidatin, Schmutz, Bestechungsgelder, Prinzipien, Praktiken, Reinigung
Fear	Gefahren, Investition, Hai, Unbekannten, Mietvertrags, Sicherheitsrisiken, Serie, Päckchen, Zielland, Bergwerk	Drohung, Geschäftstreffen, Gestalt, Personalabbau, Arbeitseinsatz, Überfälle, Absender, Heiler, Spuren, Bankkonto	Prophezeiung, Absender, Deals, Anonymität, MonatsTicket, Heiler, Fußschmerzen, Drohung, Knarren, Unterholz
Guilt	Geige, Urgroßvaters, Meeresschildkröte, Preisgeld, Sponsorenliste, Komposition, Geburtstagsjungen, Krampf, Budgets, Hotelbesitzer	Tänzerinnen, Oma, Regel, Handeln, Reisebegleiterin, Diaspora, Kaffeetasse, Einfluss, Kulturarbeit, Wartung	Oma, Opa, ExPartnerin, Reisebegleiterin, Mentee, Omas, Kaffeetasse, Neffen, Telefonanruf, Schläger
Joy	Arbeitsdruck, Uhrzeit, Winter, Fangmethoden, Kajaktrip, Geldstrafe, Angelgeräts, Unachtsamkeit, Spender, Testaments	Terminen, Halle, Tuch, Schauer, Startbahn, Verantwortliche, Schwimmkurs, backstage, Sicherheitsproblem, Hausarbeiten	Kunstmärkte, Soundsystem, Tuch, Ski, Melodie, backstage, Finanzlücke, Beine, Korrekturplan, Sicherheitsproblem
Pride	Gemeinschaft, Projektteams, Freiwilliger, Sohn, Wanderern, Höhle, Rückschlag, MarathonTraining, Testfahrt, Okay	Schützling, Neffe, Beweis, Integration, Arbeitslosigkeit, Version, Gartenmarkt, Schnorchel, Bocciaplätzen, Fernschuss	Ausdauer, Schützling, Geschwisterkind, Neffe, Lernmethode, Restauration, Tunnels, Unterrichtsmethode, Bocciaplätzen, Fernschuss
Relief	Gewitterregen, Verluste, Beobachtung, Umleitung, Strafe, Flop, Verlegenheit, Einflüssen, Leck, Blitz	Druckerei, Soundsystem, Sonnenbrille, Tradition, Vorsichtsmaßnahmen, Bühnendekoration, Symptomen, Kontrahent, Hitze, Abwasserrohr	Sandbau, Koordination, Druckerei, Tanzlied, Sonnenbrille, Hauptkünstler, Forschungsmethode, Deutschland, Fieber, Kontrahent
Sadness	Falle, Exposition, Schlittschuhe, Skiführer, Kameraden, Opa, Tauchabenteuer, Vogelpfleger, Ladenfenster, Zeltkonstruktion	Opa, Oma, Grab, Opas, ExPartnerin, Tod, Mentors, Trennungsmitteilung, Kühlschrank, Lieblingsgericht	Opa, Oma, Grab, Tänzerinnen, Omas, Kochbuch, Gedenkgang, Gentrifizierung, Opas, Auflösung
Shame	Hündin, Leichtigkeit, Korb, Gläser, Zeltstadt, Riss, Hans, Feigling, Sicherungsmaßnahmen, Vorkenntnisse	Affäre, Folien, Blogbeitrag, Ehre, Fastenkur, Katze, Date, Alkohol, Aktionen, Unwissenheit	Gartenparty, Defizite, Rückenprobleme, Fastenkur, Drohung, Anführer, Lüge, Agile, Abriss, Alkohol
Surprise	Kleinigkeit, Kritiken, Besucherzahl, Currywurst, Klarheit, Vorsprung, Bewerbungsfrist, Tagesordnung, Projektmanagement, Mitarbeiterentwicklung	Paddel, Kindheitserinnerungen, Erblasser, Adoptionsagentur, Max, Symbole, Partei, Angelplatz, Ornithologe, Geeignetes	Kontinuität, Erblasser, Menschenrechtsverletzungen, Schaben, Hauptinterpret, Operationssaal, Volunteers, Werbeanzeige, Paddel, Max
Trust	Lebensumständen, Regenschirm, Handwerker, Mitgefühl, Praxis, Studie, Instituts, Streckenänderungen, Wettbewerbsort, Verwaltung	Anna, Fremde, Einheimische, Projektmanagerin, Instructor, Wildnis, Bergführer, Fuße, Anbieter, Erzählabende	Anna, Instructor, Alex, Projektmanagerin, Mentors, Organisatorin, Berghütte, Lehrers, Umsatzrückgang, Formfehlers
No-Emotion	Wärme, Seils, Dartpfeilwerfer, Kopfhörer, Favoritenteam, Fährverbindung, Campen, Sprungturm, Warteschlangen, Idole	Festaktes, Einsatzort, Tauchen, Formalität, Eingriff, Bewegungen, Emotionen, Riss, Forscherin, RoutineÜberprüfung	Agentur, FitnessApp, Institut, Pflichtgefühl, Festwagen, Begründung, streit, Organizer, Forscherin, Protokoll

Table 16: Top 10 tf-idf nouns in event chains generated by different methods based on event Set 3.

Subset	Baseline	PC	PCR
Wut	Enkel, Partyplanerin, Prüfer, Zustimmung, Stressbewältigung, Fleisch, Begründung, Büros, Horoskops, Lieblingsbluse	Bau, Begründung, Punktzahl, Einbeziehung, Versicherungsvertreter, Verursacher, Gesundheitsinitiative, Schwangere, Kur, Traumstelle	Gleichgültigkeit, Petition, Eiscreme, Punktzahl, Umweltschützers, Teamsitzung, Versicherungsvertreter, Verursacher, Gesundheitsinitiative, Schwangere
Langeweile	Dutzende, Tatendrang, Kursabschluss, Lieblingskleidung, Schuljahr, Fußballspiel, Marathonlauf, Lieblingsserie, Überraschungspartys, Gleichgültigkeit	Leistungssportler, Dach, Kletterparcours, Abenteuerurlaub, Verleihung, Retreat, Haustier, Mittag, Stellenbeschreibungen, Familienaktivitäten	Kletterparcours, Schneemassen, Erzählabend, Abenteuerurlaub, Leistungssportler, Retreat, Karriereperspektiven, Stellenbeschreibungen, Setlist, Komponisten
Ekel	Ölfass, Literaturleser, Substanz, Fußball, Frettchen, Verunreinigung, Pestizid, Artikeln, Märkten, Psychologen	Skandale, Hygiene, Beweise, Hygienestandards, Praktiken, Korruption, Geschäftspraktiken, Unregelmäßigkeiten, Betrug, Verschmutzung	Geschäftspraktiken, Korruption, Hygienestandards, Skandale, Beweise, Hygiene, Ausbeutung, Insekten, Vorwürfe, Maden
Angst	Heißluftballons, Tänzerinnen, Horrorfilm, Einbrüche, Fotoshoot, Elfmeter, Tauchzertifizierung, Moderatoren, Anrufer, Kriminalitätsraten	Wilderer, Absender, Athletin, Unruhen, Hautpflegeserie, KI, Vorsicht, Nachrichtenmeldung, Wanderpartner, Artefakt	Artefakts, Athletin, Allergie, Wilderer, Phänomenen, Waldbrand, Familiengeheimnisse, Überfällen, Diebstählen, Unruhen
Schuld	Bekanntschaft, Gesprächstages, Schwägerin, Wohnwagen, Umweltbranche, Helm, Geldbeutel, Nichte, Gehaltserhöhung, Grill	Gedicht, RollerSkates, Ex, Serum, Warnschilder, Geheimnisses, Sicherheitsbeauftragte, Handlungen, Bräutigam, Schlafsäcke	Versuchung, RollerSkates, Medikation, Gedicht, Fokussierung, Karpfenpopulation, Kunsthandwerks, Organisationsgruppe, Museumsmanagement, Serum
Freude	Führungsteam, Hartarbeit, Fahrunterricht, Gartenhaus, Reitstundenlehrer, Reagenz, Fähnchen, Mitarbeiterunzufriedenheit, Rekrutierungsstrategien, Familientreffend	Enkelin, Monatskarte, Familientag, Ehepaar, Diskussionskultur, Gartenarbeit, Doktorvater, Hochzeitstag, Lauftraining, Nationalparks	Enkelin, Monatskarte, Lieblingsband, Lieblingsspieler, Zeltplatz, Lauftraining, Familientag, Street, Nationalparks, Himmels
Stolz	Bau, App, Sandkasten, Schlüsselpersonen, Schmuckkasten, Restaurierung, Sohn, Herstellung, Wellnesstuch, Technikzelts	Prototyp, Schatz, Rekrutierungskampagne, Bestätigungen, Rettungsteams, Mediation, Standort, Lummerkasten, Trail, Schneebesens	Prototyp, Improvisation, Rekrutierungskampagne, Bestätigungen, Grenzpassage, Mediation, Zeltlagers, Verletzten, Schneebesens, Perfektionierung
Erleichterung	Systemen, Grundbedürfnisse, Überschwemmungen, Fotobox, Ausweise, Schließung, Horizont, Aktion, Arbeitsplätze, Stocken	Segel, Trophäe, Gegenstand, Campingausflugs, Fristen, Projektteilnehmer, Thermometer, Pannen, Schlimmste, Max	Trophäe, Segel, Strafstoß, Berufswechsel, Campingausflugs, Zwangsräumung, Thermometer, Gegenstand, Pannen, Max
Traurigkeit	Trauer, Grab, Weggang, Tanzfest, Retreat, Pfleger, Grill, Kochausbildung, Sammeln, Luna	Alex, Wandmalereien, Titel, Maschine, Abschiedsfeier, ExFreund, Fitnesspartner, Abschiedsausstellung, Schmuckstück, Teamleader	Alex, Großvater, ExFreund, Maschine, ExPartner, ExFreundin, Abschiedsfeier, Lachens, Fitnesspartner, Wandmalereien
Scham	Lungen, Held, Kunstwettbewerb, Challenge, Aufschrift, Markttages, YogaStunde, Angelwettbewerb, Haken, Benzin	Fehlverhalten, Steak, Marktforschungskampagne, Shooting, Theatergruppe, Mahnung, Witz, Alkohol, Instrument, Wahrheit	Lüge, Steak, Verhaltens, Alkohol, Zelts, Marktforschungskampagne, Theatergruppe, Produzenten, Fehlverhalten, Behauptungen
Überraschung	Ankündigungen, Abteilungsleiterin, Versicherungspolicy, Nachbarin, Schneeschuhwandern, Torpedoschlauch, Annahme, Rennort, Tokio, Laternenpfahl	Reitstiefel, EliteSchule, Starre, Gemeinschaftsraums, Schneeschuhe, Band, Mittag, Bar, Telefonats, Surfausflug	Sternen, Reitstiefel, EliteSchule, Schneeschuhe, Reisepass, Band, Highlights, Therapien, Ehrengäste, Bar
Vertrauen	Ausstellungsleitung, Ashram, Reformierung, Fotobuch, Sprachkenntnisse, Außenseiters, Hartarbeit, Beobachtungsplatzes, Weggabelung, Skitour	Beratungsgespräch, Mentorin, Webseite, Gründlichkeit, Kommunikations, Neubeginn, Vogelart, Surflehrer, Bauch, Rudertour	Eiche, Beratungsgespräch, Schlittenführer, Webseite, Mentorin, Navigation, Erkenntnissen, Gegenseite, Einheimische, Kurslehrer
keine_ Emotion	Gelassenheit, Kasse, Verlangen, Ritual, Siege, Kletteraufstieg, Priorität, Handbuchs, Bürogebäudes, Personalabteilungsleiterin	Rollern, HobbyAbend, Sportgeräte, Auswahlkomitee, Meditationssession, Faktoren, Abschnitte, Flamme, Reflektion, Elektroautos	Meditationssession, HobbyAbend, Sportgeräte, JobShadowing, Patientin, Abstieg, ordnung, Auswahlkomitee, Vertrautheit, Stürme

Table 17: Top 10 tf-idf nouns in event chains generated by different methods based on event Set 4.