Department of Information Engineering and Computer Science

Master's Degree in
Artificial Intelligence Systems

Final Dissertation

Automatic Narrative Elicitation with Large Language Models

Supervisor
Prof. Giuseppe Riccardi

Co-Supervisor
Gabriel Roccabruna

Student
Michele Yin

Academic year 2022/2023

Acknowledgements

... to my kind supporting family, my father 尹利平, mother 郭玲琴, and my sister Silvia ... to the supervisor Prof. Giuseppe Riccardi and co-supervisor Gabriel Roccabruna ... to all my friends who had supported me during this path ... special remembrance goes to my dear friend, Samuele Conti, whose memory and will inspired me ...

... in the process of completing this thesis and I am struck by the loss of Filippo Momesso. I am reminded of the critical importance of mental health. Filippo's tragic decision underlines the urgency of addressing mental health challenges in our society ...

Contents

List of Figures

List of Tables

Abstract

Narratives serve as a means of communication and are used to tell stories, share information, express ideas, or entertain. However, recounting narratives can be challenging due to memory limitations, selective recall, information loss over time, and many other problems. In fields such as psychology, sociology, anthropology, and qualitative research, elicitation techniques are often used to collect narratives or stories from individuals. It typically involves cueing individuals to share personal stories, anecdotes, or narratives related to a specific topic, event, or experience. In psychology, elicitation techniques may be employed to improve communication and aid problem-solving. However, narrative interviews as elicitation face some challenges, as they require a nuanced approach that involves empathetic listening and allowing the storyteller to shape their own account.

This thesis investigates the application of Large Language Models in the newly defined task of *Automatic Narrative Elicitation*. In this task, each narrative is recounted by the narrator and the model is given the role of interviewer that generates questions that elicit the continuation of an interrupted narrative, prolonging the narration act from the point of interruption. We also present a novel crowdsourced dataset in the Italian language designed for the same task. The model results are compared with the crowdsourced dataset using both automatic metrics and human evaluation to ensure a comprehensive and correct understanding of the capabilities of the models that were tested. Our results show that adequately prompted LLMs can reach the performance of human annotators for the task of Automatic Narrative Elicitation.

1 Introduction

Personal narratives are a form of storytelling, representing the recollections of events or connected sequences of events in which the narrator has played an active or passive role [54]. These narratives, often intimate and reflective, find their expression in various mediums, from handwritten diaries to digital travelogues, encompassing both speech and text. The power of personal narratives lies in their ability to convey individual experiences, emotions, and perspectives, making them a rich source of insight into the human condition [1, 10, 8] in particular for the domain of mental health [37].

The act of narration, however, is not a straightforward task. The narrator may experience issues during their narration, interrupting themselves because they may not know how to continue their narrative or they might not have a clear narrative flow in mind [38]. Other issues involve difficulties in the narrator of sharing their stories due them being uncomfortable or due to the presence of sensitive topics in it [52]. Narrative interviewing can be viewed as a way to mitigate these issues, with the interviewer inducing the continuation of the narration through questions. However, this task goes beyond merely asking questions; it requires a nuanced approach that involves empathetic listening and allowing the storyteller to shape their own account [27]. This delicate interplay between interviewer and interviewee underscores the complexity of capturing authentic personal narratives.

In recent years, Large Language Models (LLMs) have achieved remarkable success in a wide range of applications and domains, including natural language understanding and generation [42, 55, 12]. Their capacity to comprehend and generate human-like text [53] has us posit whether LLMs can effectively elicit the continuation of personal narratives.

This thesis investigates the performance of LLMs in the novel task of *Automatic Narrative Elicitation* (ANE). The aim is to investigate whether these models can be prompted to engage in interactions that lead to the correct elicitation of continuation of personal narratives, following a structure similar to the narrative interview. This means that these elicitations should be questions, focusing on topics mentioned in the narrative and conveying the feeling of active and interested listening to the narrator.

In this task, each narrative is recounted by the narrator and the model is given the role of interviewer that generates questions that elicit the continuation of an interrupted narrative, prolonging the narration act from the point of interruption. To be effective, these questions have to to be short and focused on an event or topic presented in the narrative, conveying empathy and most importantly, conveying the feeling that the model is actively listening to the narrative, because this is a key requirement in an effective narrative interviewing [27].

To train and test deep learning models on this task, we crowdsourced a set of eliciting questions for each of the personal narratives in the CoAdapt corpus. This corpus contains 481 personal narratives in the Italian language from 45 subjects, who were employees with stress undertaking the Cognitive Behavioural therapy [16]. Despite the growing number of LLMs, only a few of them are suitable for the Italian language. To skim them, we designed an Italian language test to easily determine which models were suitable and which were not. The few models that could answer in Italian were subjected to the *Story Cloze Test* [34] because this task requires a correct understanding of commonsense knowledge to predict the story closure correctly. This commonsense inference and reasoning ability is a proxy for the key listening requirement present in ANE.

Finally, we assessed the select LLMs on the ANE task.Although the results of automatic evaluation suggest that the models generate different questions from the crowdsourced ones, the results of human evaluation suggest that with some prompts, specific LLMs have the potential to attain performance levels on par with those of crowdworkers in the ANE task.

The thesis is organized as follows:

- Chapter 1: Introduction, which is this chapter.
- Chapter 2: Literature Review of the related topics.
- Chapter 3: Methodology. This chapter contains a description of the presented corpus, the LLMs selection and

prompting procedure.

- Chapter 4: Evaluation. This chapter contains the evaluation of the results of LLMs, with both automatic and human evaluation metrics
- Chapter 5: Conclusion. This chapter provides a summary of the dissertation and proposes some possible future works.

2 Literature review

This section consists of a literature review of the concepts of personal narrative, narrative interview, large language models, and prompt engineering.

2.1 Personal Narrative

Personal narratives are a form of storytelling, representing the recollections of events or connected sequences of events in which the narrator has played an active or passive role [54]. These narratives, often intimate and reflective, find their expression in various mediums, from handwritten diaries to digital travelogues, encompassing both speech and text. Personal narratives serve as a means of communication and are used to tell stories, share information, express ideas, or entertain. They can be used to convey personal experiences, beliefs, values, and perspectives. Personal narratives are often used in fields such as psychology, sociology, anthropology, and qualitative research to gather in-depth qualitative data and gain insights into individuals' perspectives, beliefs, and lived experiences [37, 14]. They can help researchers understand how people construct and interpret their personal narratives and the meanings they attribute to various events and experiences. For instance, personal narratives may be employed in psychology to improve communication and aid problem-solving [27].

2.2 Narrative Interview

A narrative interview is a qualitative research method to gather in-depth information about an individual's experiences, stories, and life events. It is a form of semi-structured or open-ended interview in which the interviewer encourages the interviewee to share their life narratives, anecdotes, and stories. A narrative interview aims to capture rich and detailed accounts of the interviewee's experiences, emotions, perspectives, and the context of these events by asking interviewees questions designed to have the participant respond in a narrative [25].

There are a few caveats in the narrative interview, as it requires the interviewee to be able to recall and recount their experiences coherently and meaningfully. This process can be challenging for some individuals, especially if they have difficulty recalling events. It can also be difficult for interviewers to elicit continuations of narratives from interviewees who feel uncomfortable sharing personal information or having difficulty expressing themselves verbally. [52]

The interviewer also requires some skill in eliciting continuations of narratives from the interviewee. They must be able to ask questions that encourage the interviewee to share their experiences and stories coherently and meaningfully. They must also be able to listen attentively and ask follow-up questions to clarify or expand on the interviewee's responses. They must also convey active interest in the narration or events mentioned in the narrative, providing empathy and understanding. [18]

2.3 Large Language Models

In recent times, there have been notable developments in the field of natural language processing through the introduction of pre-trained language models [47]. These models involve pre-training transformer-based architectures [56] on extensive text corpora and have demonstrated remarkable capabilities across various NLP tasks. As researchers delved into the potential for performance enhancement by scaling upwards the size of their models, they found interesting results. As the scale of these models exceeded certain thresholds, they exhibited significant performance improvements and showcased unique abilities not observed in smaller language models. The research community referred to these models as large language models among the pre-trained language models of substantial proportions to distinguish these models based on their parameter size. [67, 57]

The exploration of LLMs has seen substantial advancements driven by both academic and industrial research efforts. One noteworthy milestone in this progress is the introduction of ChatGPT [40], which has garnered widespread attention and interest from society due to its capabilities and ease of use. LLMs in general, are found to perform well for a large variety of tasks, exhibiting human-like performances and by some are even considered to be the next step in the evolution of artificial intelligence [13].

However, they also face a few issues. Notably, one is the problem called hallucinations. Hallucinations in large language models refer to instances where the model generates text or responses that are not grounded in factual information or are not based on the input provided. These hallucinations can manifest as the model generating fictional or inaccurate information that appears true, coherent, or contextually relevant, but in reality, it is incorrect. The models are often very confident in the correctness of their responses, making hallucinations difficult to detect, posing a challenge to the responsible use of these models [4, 6]. Hallucinations are a known challenge in developing and deploying large language models. They can arise due to the capacity of the model to generate creative and contextually coherent text based on patterns learned from the training data, even if those patterns do not reflect reality or are not appropriate for the given context. Addressing and mitigating hallucinations is an ongoing area of research and development [21, 36, 45]. Another important issue concerns their large size. The sheer number of parameters present can seen as a significant barrier to their use, increasing the cost of access and use, as only high-end expensive GPUs can run many of the available models. For example, ChatGPT 3.5 has 175B parameters [62], and a rough math of storing the parameters as 16 bits precision estimates the requirement of running ChatGPT 3.5 to 350GB of GPU memory. Although there are many developments in crafting performing models with smaller memory requirements, they have some tradeoffs. Many smaller models come in sizes of just 7B or 13B parameters and are able to run on a single GPU with 16GB or 25GB of memory. However, they do not perform as well as the larger models and are not as capable [55, 5, 32, 2].

2.4 Prompt Engineering

Prompt engineering refers to designing and formulating effective input prompts or instructions to elicit desired responses from text-input models. In the case of large language models, these responses are text-based, but for text-to-image generation models, their responses are images. It involves crafting input text so that the model produces outputs that align with the user's intent and the required task. Prompt engineering is essential for improving the models' large language usability and reliability, especially when used in various applications, such as text generation, question answering, or language translation [60, 70, 43, 48, 69].

One of the most important key aspects in designing prompts is the number of shots or examples that are given to the model. Usually, it is generally accepted that more examples help the model perform better [12, 48, 3, 59], although some recent findings suggest that LLMs can perform well with 0 examples as well [28]. Other key aspects that prompts for large language models should include are: clarity and specificity, context setting, formatting and structure. Interestingly, chain-of-thought and think-step-by-step are two techniques that are found to improve the performances of the model in logical reasoning tasks by asking the model to perform the task in a step-by-step procedure [58, 23]. Role-giving is also found to be an important aspect in prompting, as it helps the model better understand the task and generate more relevant responses. Providing an appropriate role can help the model generate responses that are more relevant to the domain [3].

There are also a few ethical considerations when designing prompts; it is important to consider ethical guidelines and avoid prompts that may lead to harmful, biased, or inappropriate outputs. Care should be taken to ensure responsible and safe use of language models [29, 26].

3 Methodology

This section encompasses the methodology employed throughout this thesis. This section includes a definition of the new task of *Automatic Narrative Elicitation* (ANE) and of its challenges, followed by the curation of a new corpus designed for this task. This corpus is composed of narratives for which eliciting questions were crowdsourced. Then, these human-sourced eliciting questions are used as reference points against eliciting questions sourced by prompting LLMs.

3.1 Automatic Narrative Elicitation

We define the task of *Automatic Narrative Elicitation* (ANE) as the task of prompting specific questions related to a given personal narrative, with the goal of continuing the current narrative, exploring events or topics mentioned in it. The model is given the role as the interviewer, and its responses should be inherent to the topics expressed in the narrative, with a focus on conveying empathy when necessitated.

In particular, using LLMs for ANE has some critical requirements for their abilities. For one, it requires LLMs to understand the concept of a conversation between two people, with the LLMs acting as the interviewer. Furthermore, LLMs are also required to understand correctly the events described in the narrative, posing questions on relevant topics mentioned in the narrative or related topics without significantly altering the flow of the narration. Moreover, these narratives may not always use correct grammar or syntax, they may lack a clear and fluent description of the events, and it is up to the interviewer, in this case the LLMs, to understand and piece together the information. Lastly, the questions should be empathetic for narratives of a sorrowful or joyful nature, which requires an understanding of empathy and the ability to convey empathetic responses.

3.2 Corpus

Given the absence of a pre-existing dataset tailored for the ANE task , we decided to design and perform a data collection to produce a new dataset. To do this, we started from a dataset of personal narratives, CoAdapt [16]. The CoAdapt dataset was initially collected within the framework of a psychological study focusing on mental health. The inception of the dataset involved soliciting users' responses daily, with a fixed set of questions presented to them. Data for *Emotion Carriers* and *Valence* was collected. Emotion Carriers are people, objects or events that indirectly convey emotions [54], while Valence represents the level of sadness or joy, measured on a scale from -2 to +2, of functional units of the narrative [39, 49, 50].

An example is shown in Table 3.1 The dataset in question was collected in the Italian language, with the active participation of native Italian speakers as data contributors. Nonetheless, care should be put in the dataset curation process, due to the intrinsic variability of human responses to such questions. There were instances when the user did not answer some of the questions or answered to multiple questions simultaneously. As shown in the previous table, simply concatenating the answers to each question together to obtain a fluent narrative is ineffective. Doing so would create incoherent narratives with no clear flow and missing linguistic conjunctions. This is due to the fact that the questions were not designed to elicit a continuation of the narrative but rather to elicit a response to those specific questions related to mental health. For this reason, the dataset was manually reviewed and refined to ensure a coherent narrative flow. This process involved the addition of conjunctions and linguistic modifications, ensuring a seamless and cohesive narrative flow. Additionally, the dataset underwent an anonymisation procedure, which entailed the removal of any references to specific locations or individuals, replacing them with randomly selected alternatives. Narratives deemed excessively brief were removed, while longer narratives were strategically divided into multiple narratives. Example of the procedure is provided in Table 3.2.

The result was a dataset tailored for the specific ANE task. Notably, this dataset comprises of a total of 476 narratives, thoughtfully divided into two subsets: 419 narratives for the training set and 57 narratives for the test set. An analysis of tokens was conducted using the Spacy tokeniser to gain further insights into the linguistic properties of the dataset. Punctuation marks, stop words, and numerical digits were excluded from this analysis, revealing the presence of 3,376 unique tokens within the dataset. Each narrative prompt spans 49.5 tokens on average, with a

Coadapt Original Data			
Question	**Answer**	**Emotion Carrier**	**Valence**
initial note	Serenità coi fiori	~	+1
cbt response a	Domenica nel mio giardino	~	0
cbt response b	Sarebbe bello poter avere un profumo simile a quello delle viole o dell' iris	poter avere, un profumo	+1
cbt response c	Occupandomi dei miei fiori ho sentito una sensazione piacevole data dal profumo delle viole e dal sole che leggero accarezzava la pelle. Ho provato Felicità nelle parti del corpo: Testa sensazione di inebriamento.	ho sentito, profumo, vuole, sole, accarezzava	+1

Table 3.1: An example of original data and the respective emotion carriers and valences. On the first column, the division in questions, and on the second, the respective textual answers. Emotion carriers, reported in the third column convey narrators' emotions in the text. The last column reports the values of valence which represents in a range from -2 to +2 the level of sadness or joy of the respective highlighted sections.

Example of CoAdapt Data and Revision		
Coadapt Original Data		**Revised Narrative**
Question	**Answer**	Domenica nel mio giardino, occupandomi dei miei fiori ho sentito una sensazione piacevole data dal profumo delle viole e dal sole che leggero accarezzava la pelle. Sarebbe bello poter avere un profumo simile a quello delle viole o dell' iris.
initial note	Serenità coi fiori	
cbt response a	Domenica nel mio giardino	
cbt response b	Sarebbe bello poter avere un profumo simile a quello delle viole o dell' iris	
cbt response c	Occupandomi dei miei fiori ho sentito una sensazione piacevole data dal profumo delle viole e dal sole che leggero accarezzava la pelle. Ho provato Felicità nelle parti del corpo: Testa sensazione di inebriamento.	

Table 3.2: Example of original data and revision that was applied to it. The original data is composed of answers to a set of 4 questions, which are not structured in a narrative format. A simple concatenation of all the text present in the data column would not suffice to obtain a narrative. Therefore manual review was applied in order to get a coherent and fluid narrative, of which an example is reported in the rightmost column.

Statistics of the narratives			
Statistic	**Train Set**	**Test Set**	**Overall Set**
Number of narratives	419	57	476
Average narrative length	50.4	44.4	49.5
Standard deviation on narrative length	38.9	31.3	37.9
Number of unique words	3225	629	3378

Table 3.3: Report of some statistics computed from narratives present in the dataset. Reported are the number of narratives, average narrative length and standard deviation in number of tokens and the number of unique words for the complete dataset, as well as the splits of train and test sets. Tokenization was done through Spacy

notably high standard deviation of 37.8. These statistics can be found in Table 3.3.

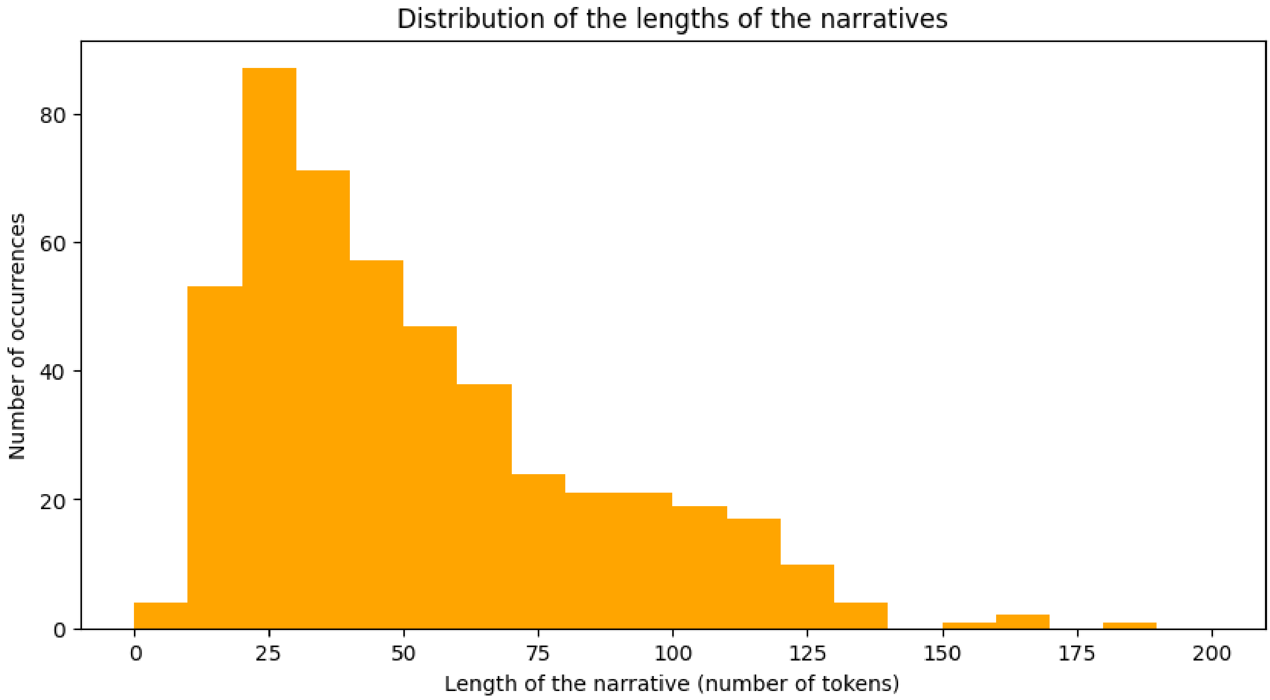


Figure 3.1: Histogram of the distribution of the lengths of the narratives. There are a few narratives that are very short, while others are very long, but the majority of the narratives are between 20 and 70 tokens, following a bell shaped curve.

The token count in some narratives is as low as 5, while others extend well beyond 200 tokens, attesting to the diversity of the dataset in terms of textual length. The histogram of the narrative lengths is reported in Figure 3.1.

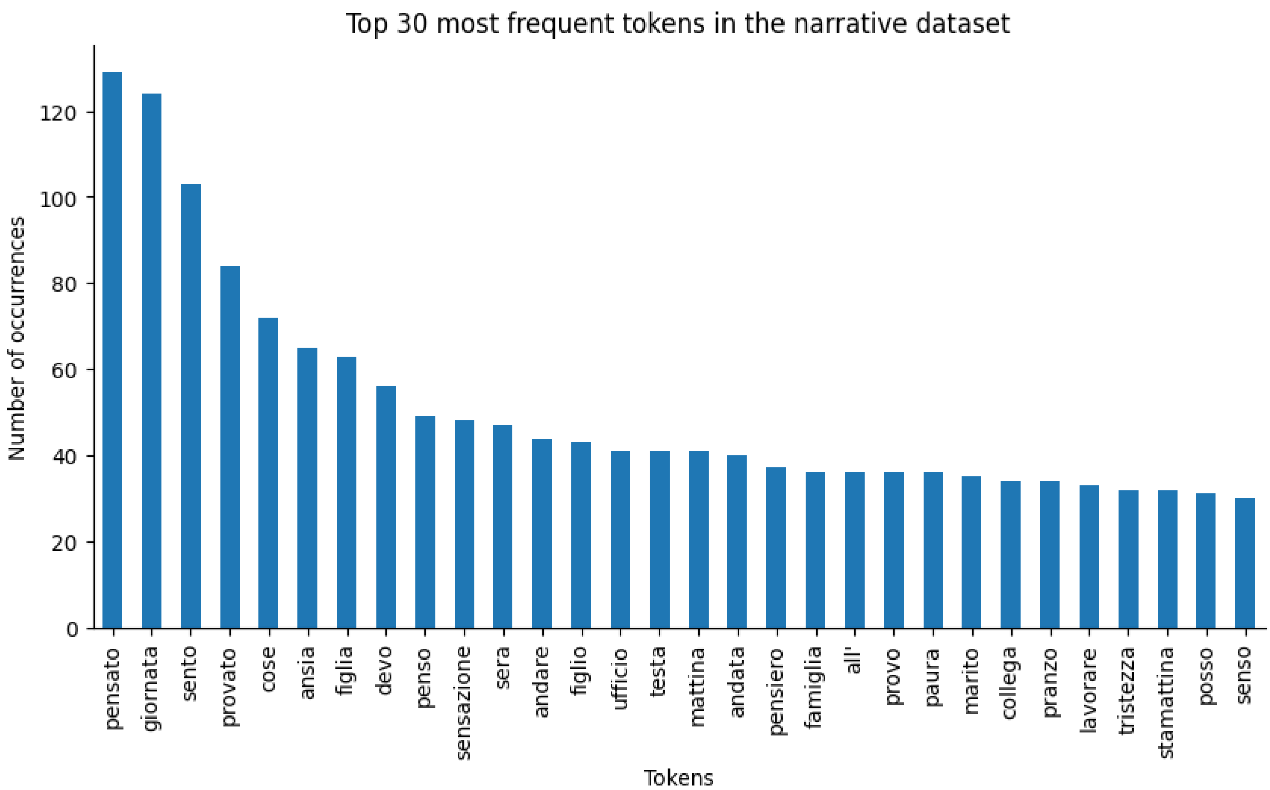


Figure 3.2: Bar chart of the top 30 most frequent tokens in the dataset. Notice how the most frequent tokens are related to anxiety, day, feelings, work, and family. This aligns with the thematic focus on mental health of the dataset.

Besides those statistics, a bar chart of the top most frequent 30 tokens is illustrated in Figure 3.2. As depicted in the figure, some of the most frequently occurring tokens align seamlessly with the inherent thematic focus on the domains of mental health, work, and family of the dataset.

3.3 Crowdsourcing

In order to complete the corpus, we required human sourced elicitations for continuations of the personal narratives presented in the dataset. We decided to exploit crowdsourcing, designing an experimental setup to instruct crowdworkers on how to induce the continuation of the narrative. This is followed by the data collection and an statistical analysis of the crowdsourced eliting questions.

3.3.1 Guidelines design

The formulation of effective guidelines was a fundamental requirement. This step necessitated an iterative process involving the creation of drafts, subsequent reviews, and pilot tests of guidelines. Pilot test were used to gauge the effectiveness of the guidelines in directing workers to generating correct eliciting questions for the continuation of personal narratives. This iterative cycle was repeated until the guidelines reached a level of satisfaction.

Example of how highlighted text is used to convey valence information

Example	Narrative	Valence values
1	Ciao, tutto bene molto lavoro il questi ultimi giorni.	+1
2	Ritornata dal lavoro mia figlia mi dice di aver chiamato il medico perché ha dei dolori alla testa non ha il senso dell'olfatto e del gusto per cui ci siamo un po' allarmati.	-1
3	Oggi mi sono dedicata al giardino e all'orto. Sono stanchissima fisicamente ma rilassata mentalmente. Fuori in giardino ho fatto tutte le cose che prima faceva mio marito. Ho sentito quasi fisicamente la sua presenza e la cosa mi ha rilassato	+1

Table 3.4: Three examples of narratives with highlighted text that represents valence values, which are reported on the right. The sections highlighted in red represent negative values, while in the ones highlighted in green represent positive values. Neutral values corresponding to 0 are not highlighted and not reported.

After finalising the guidelines, the next step involved translating them into a custom web-based user interface (UI) for the data collection process. One key aspect of our guidelines was the inclusion of the valence values from the CoAdapt dataset. As previously mentioned, effective elicitation requires empathy, particularly for sorrowful events. In order to facilitate the crowdworkers in pivoting their questions to emotionally charged events, the valence values were highlighted in the text, with the recommendation to focus on those parts of the narrative. In order to represent effectively to the crowdworkers the positive and negative emotions expressed by the valence, it was decided to highlight negative valences in red and positive valences in green. An illustrative example can be seen in Table 3.4. These colours were chosen as it is universally accepted that green stands for positive and red for negative. Furthermore, to limit the crowdworkers' cognitive workload, the ECs have not been added.

This task places a large emphasis on the examples included in the guidelines because, during a pilot test, it was found that many crowdworkers try to complete the task as quickly as possible, giving only a light read to the guidelines and skipping directly to the examples. Table 3.6 reports examples of good and bad elicitating questions for a given narrative that were given to the crowdworkers.

A brief recap of the desired properties of the eliciting questions is reported in Table 3.5. The table reports a few scenarios, such as requests for personal opinions or hypothetical conditions, which should be avoided because, in the internal testing, it was found that those types of questions may induce the narrator to start a new narrative with a different set of events. This is not desirable as the goal of those eliciting questions is to explore the current narrative and not to start a new one.

3.3.2 User Interface (UI) Design for Crowdsourcing

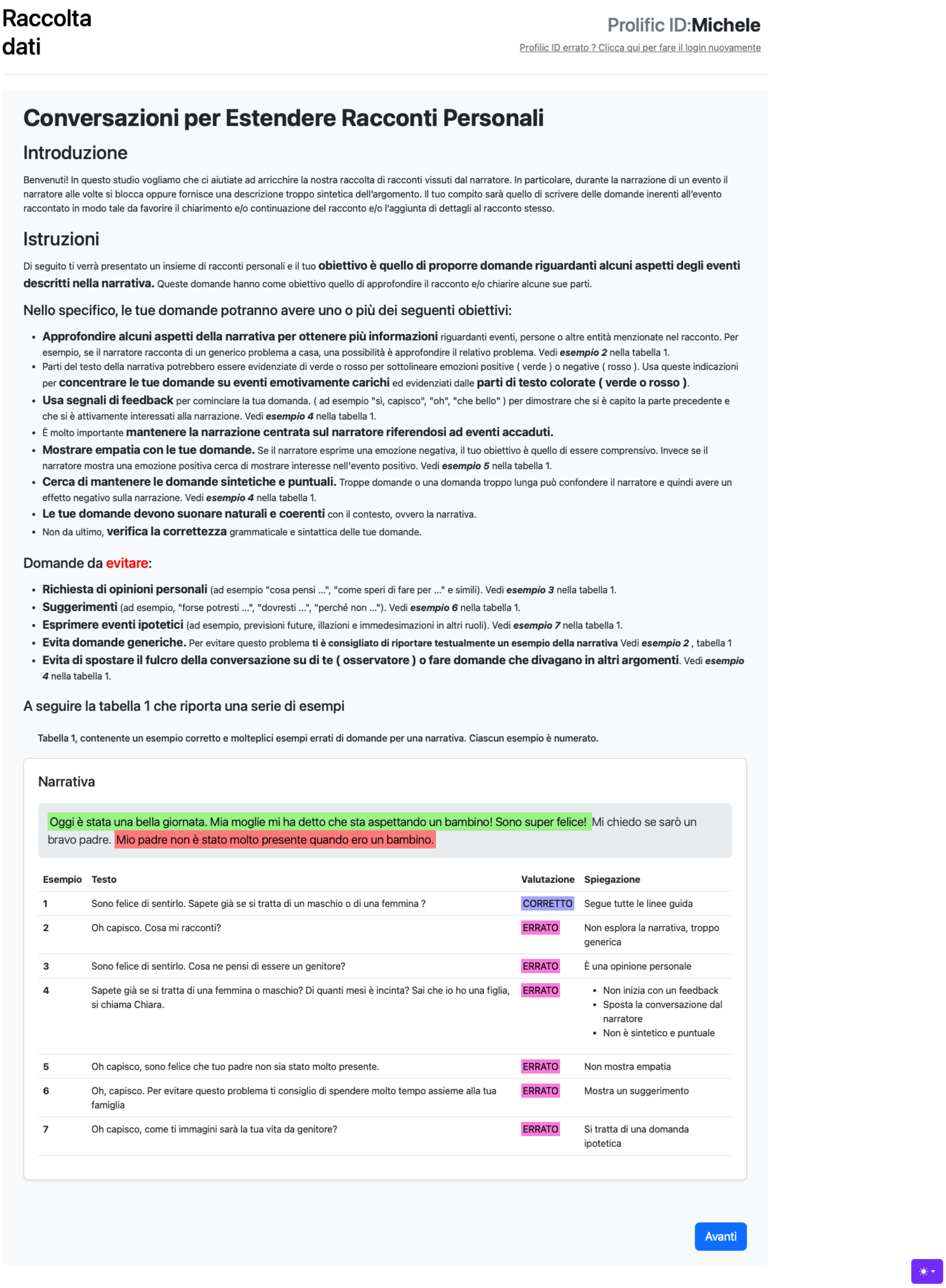
For the UI, the selection of the design framework was founded upon Bootstrap [11]. Bootstrap was chosen due to its user-friendly nature, ease of development, and contemporary design elements. Testing was conducted on multiple

Desired and Undesired properties of the elicitating questions from the crowdworkers			
Number	**Desired properties**	**Number**	**Undesired properties**
1	Suggested, but not enforced, focused on the highlighted portions of the narratives, i.e. parts of the narratives with valence.	**1**	Questions on personal opinions.
2	Containing feedback signals, such as *"Capisco"* or *"Oh, che bello"* and other signs of active listening.	**2**	Suggestions
3	Centered around the narrator, i.e. should not move away from the narrator the focus of the story.	**3**	Hypothetical questions or scenarios.
4	Containing explicit references to events that happened in the narrative.	**4**	Questions that move the focus of the conversation away from the narrator.
5	Showing empathy to the narrator, for instance with words as *"Mi dispiace"* when sad events are mentioned.		
6	Short and on point.		
7	Correct, both grammatically and syntactically.		
8	Focused on events of the narrative.		

Table 3.5: Table reporting the desired (on the left) and undesired (on the right) properties of the eliciting questions

Examples of correct and incorrect eliciting questions			
Narrative	Oggi è stata una bella giornata. Mia moglie mi ha detto che sta aspettando un bambino! Sono super felice! Mi chiedo se sarò un bravo padre. Mio padre non è stato molto presente quando ero un bambino.		
Example	**Text**	**Evaluation**	**Explaination**
1	Sono felice di sentirlo. Sapete già se si tratta di un maschio o di una femmina ?	**CORRETTO**	Segue tutte le linee guida
2	Oh capisco. Cosa mi racconti?	**ERRATO**	Non esplora la narrativa, troppo generica

Table 3.6: Example that was provided to the users of the crowdsourcing data collection. It contains one example of narrative with one example respectively of correct and incorrect eliciting question, paired with their respective motivation.



Raccolta dati

Prolific ID:**Michele**

Prolific ID errato ? Clicca qui per fare il login nuovamente

Conversazioni per Estendere Racconti Personali

Introduzione

Benvenuti! In questo studio vogliamo che ci aiutiate ad arricchire la nostra raccolta di racconti vissuti dal narratore. In particolare, durante la narrazione di un evento il narratore alle volte si blocca oppure fornisce una descrizione troppo sintetica dell'argomento. Il tuo compito sarà quello di scrivere delle domande inerenti all'evento raccontato in modo tale da favorire il chiarimento e/o continuazione del racconto e/o l'aggiunta di dettagli al racconto stesso.

Istruzioni

Di seguito ti verrà presentato un insieme di racconti personali e il tuo **obiettivo è quello di proporre domande riguardanti alcuni aspetti degli eventi descritti nella narrativa.** Queste domande hanno come obiettivo quello di approfondire il racconto e/o chiarire alcune sue parti.

Nello specifico, le tue domande potranno avere uno o più dei seguenti obiettivi:

- **Approfondire alcuni aspetti della narrativa per ottenere più informazioni** riguardanti eventi, persone o altre entità menzionate nel racconto. Per esempio, se il narratore racconta di un generico problema a casa, una possibilità è approfondire il relativo problema. Vedi ***esempio 2*** nella tabella 1.
- Parti del testo della narrativa potrebbero essere evidenziate di verde o rosso per sottolineare emozioni positive (verde) o negative (rosso). Usa queste indicazioni per **concentrare le tue domande su eventi emotivamente carichi** ed evidenziati dalle **parti di testo colorate (verde o rosso)**.
- **Usa segnali di feedback** per cominciare la tua domanda. (ad esempio "sì, capisco", "oh", "che bello") per dimostrare che si è capito la parte precedente e che si è attivamente interessati alla narrazione. Vedi ***esempio 4*** nella tabella 1.
- È molto importante **mantenere la narrazione centrata sul narratore riferendosi ad eventi accaduti.**
- **Mostrare empatia con le tue domande.** Se il narratore esprime una emozione negativa, il tuo obiettivo è quello di essere comprensivo. Invece se il narratore mostra una emozione positiva cerca di mostrare interesse nell'evento positivo. Vedi ***esempio 5*** nella tabella 1.
- **Cerca di mantenere le domande sintetiche e puntuali.** Troppe domande o una domanda troppo lunga può confondere il narratore e quindi avere un effetto negativo sulla narrazione. Vedi ***esempio 4*** nella tabella 1.
- **Le tue domande devono suonare naturali e coerenti** con il contesto, ovvero la narrativa.
- Non da ultimo, **verifica la correttezza** grammaticale e sintattica delle tue domande.

Domande da evitare:

- **Richiesta di opinioni personali** (ad esempio "cosa pensi ...", "come speri di fare per ..." e simili). Vedi ***esempio 3*** nella tabella 1.
- **Suggerimenti** (ad esempio, "forse potresti ...", "dovresti ...", "perché non ..."). Vedi ***esempio 6*** nella tabella 1.
- **Esprimere eventi ipotetici** (ad esempio, previsioni future, illazioni e immedesimazioni in altri ruoli). Vedi ***esempio 7*** nella tabella 1.
- **Evita domande generiche.** Per evitare questo problema **ti è consigliato di riportare testualmente un esempio della narrativa** Vedi ***esempio 2***, tabella 1
- **Evita di spostare il fulcro della conversazione su di te (osservatore) o fare domande che divagano in altri argomenti**. Vedi ***esempio 4*** nella tabella 1.

A seguire la tabella 1 che riporta una serie di esempi

Tabella 1, contenente un esempio corretto e molteplici esempi errati di domande per una narrativa. Ciascun esempio è numerato.

Narrativa

Oggi è stata una bella giornata. Mia moglie mi ha detto che sta aspettando un bambino! Sono super felice! Mi chiedo se sarò un bravo padre. Mio padre non è stato molto presente quando ero un bambino.

Esempio	Testo	Valutazione	Spiegazione
1	Sono felice di sentirlo. Sapete già se si tratta di un maschio o di una femmina ?	CORRETTO	Segue tutte le linee guida
2	Oh capisco. Cosa mi racconti?	ERRATO	Non esplora la narrativa, troppo generica
3	Sono felice di sentirlo. Cosa ne pensi di essere un genitore?	ERRATO	È una opinione personale
4	Sapete già se si tratta di una femmina o maschio? Di quanti mesi è incinta? Sai che io ho una figlia, si chiama Chiara.	ERRATO	• Non inizia con un feedback • Sposta la conversazione dal narratore • Non è sintetico e puntuale
5	Oh capisco, sono felice che tuo padre non sia stato molto presente.	ERRATO	Non mostra empatia
6	Oh, capisco. Per evitare questo problema ti consiglio di spendere molto tempo assieme alla tua famiglia	ERRATO	Mostra un suggerimento
7	Oh capisco, come ti immagini sarà la tua vita da genitore?	ERRATO	Si tratta di una domanda ipotetica

Avanti

Figure 3.3: Image of the Web UI that was shown to the crowdworkers for the guidelines. This page was shown initially to the user and was available for later review with the press of a button.

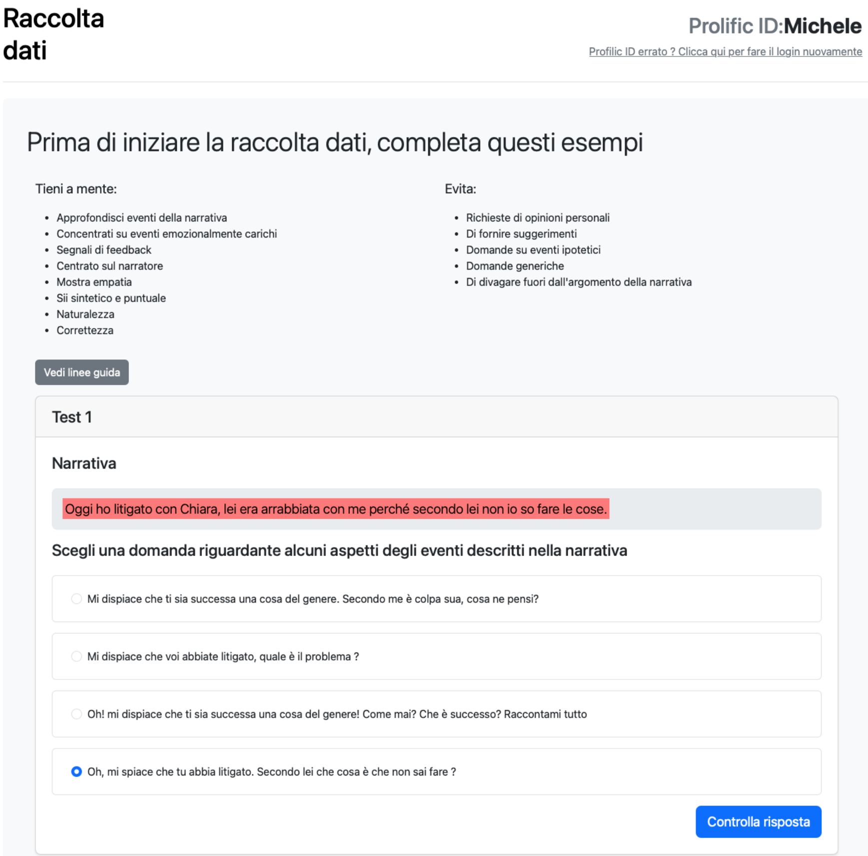


Figure 3.4: Image of the Web UI that was shown to the crowdworkers as training examples. After the guidelines, 4 examples similar to this one were assigned to the crowdworkers in order to train them.

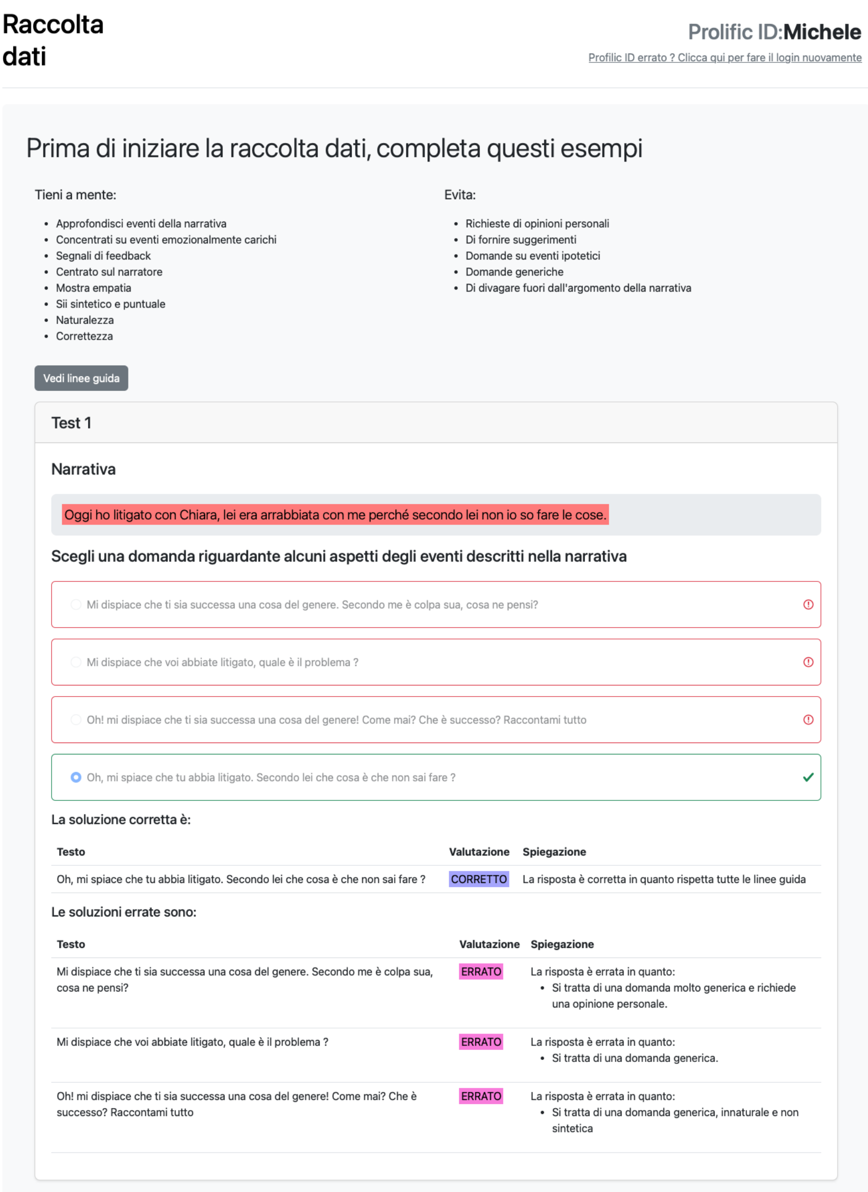


Figure 3.5: Image of the Web UI that was shown to the crowdworkers as training examples. After each user inputs an answer, an example of a correct answer this the motivation was shown to train the crowdworkers against typical mistakes.

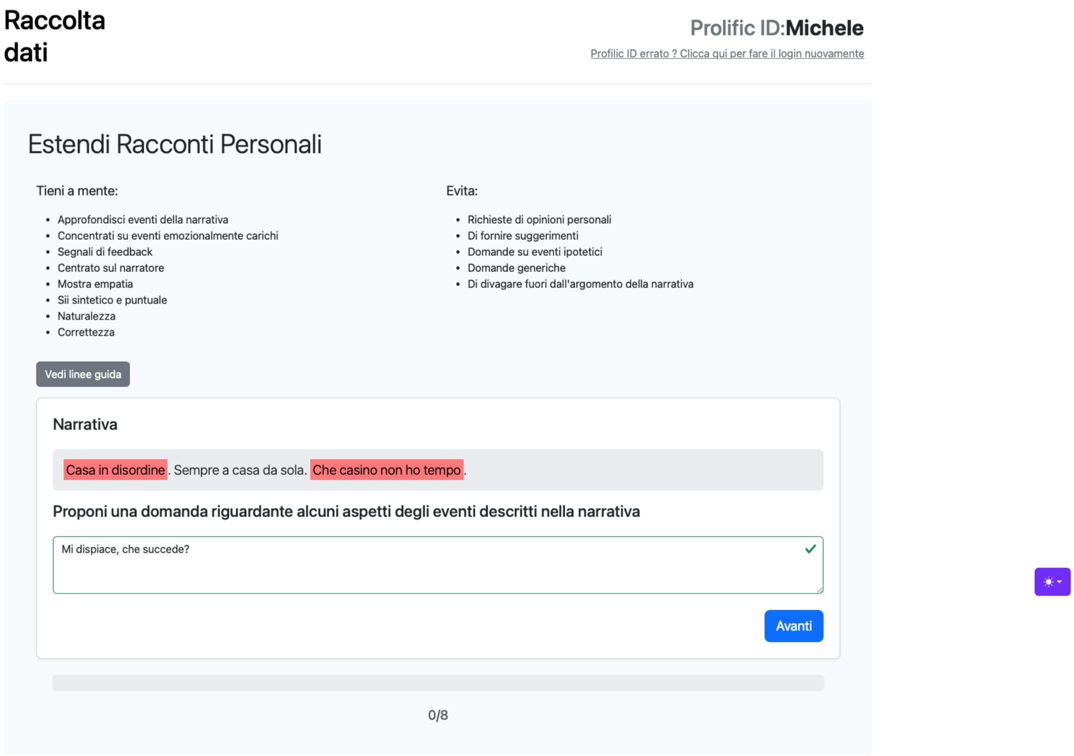


Figure 3.6: Image of the Web UI that was shown to the crowdworkers for the actual data collection process.

platforms, encompassing mobile devices, various web browsers, and distinct operating systems to ensure optimal user experience across diverse devices. A few pages of the UI are reported. In Figure 3.3 is shown the UI used to convey the guidelines, in Figure 3.4 and 3.5 are shown the examples that were used as training examples for the users. Finally, in Figure 3.6 is reported the UI used for the actual data collection.

The colours used in the palette are material [20], with modern rounded corners and no sharp edges, keeping the design as minimal as possible. Colours for the highlighted text are a pale shade of red and green, # F77A7B ( ) and # 9AF288 ( ). These colours were chosen due to their light tint and their high contrast on black text, allowing the text underneath the highlighted selection to be easily read.

3.3.3 Data Collection

Subsequently, upon the successful completion of the user interface, the data collection phase was initiated. Prolific, a reputable platform [46] for data collection, was employed for this purpose. From this platform, only native Italian speakers were selected as eligible to undertake the task, due to the fact that our corpus is written in the Italian language.

The data collection process started with an initial pilot test, verifying accuracy and identifying any potential procedural errors. Once those were fixed, the data collection was started. The data collection was conducted in multiple batches, remunerating participants at a rate of £12 per hour. To prevent participant fatigue, each batch included a range from 5 to 8 narratives, with an estimated completion time of approximately 20 minutes per batch. In order to give roughly the same amount of load to each worker, a stratified sampling approach was used on the narrative lengths, giving each worker a similar amount of words to read. At the same time, to ensure consistency and elicit agreement among narrators, the first narrative for each batch across a set of batches was intentionally kept identical for different annotators. This first narrative was also shorter in length compared to subsequent narratives. This approach was employed for the dual purpose of manually verifying consensus among narrators regarding their elicitation topics on the same narrative and also serve as a warm-up exercise for the crowdworkers. This meant that there were a few narratives with multiple eliciting questions.

Following each run, each eliciting question was individually inspected and proofread in order to reject any unreliable data, resulting in the exclusion of only one set of answers.

3.3.4 Data Analysis

Statistics of the crowdsourced eliciting questions			
Statistic	**Train Set**	**Test Set**	**Overall Set**
Number of narratives	419	57	476
Number of eliciting questions collected	510	84	594
Number of annotators	63	14	77
Number of unique words	1015	236	1110
Average eliciting question length	12.1	10.4	11.8
Standard deviation on eliciting question length	6.6	5.1	6.4
Average guidelines reading time	429 s	291 s	401 s
Standard deviation on guidelines reading time	336 s	126 s	311 s
Average narrative elicitating question time	97 s	102 s	98 s
Standard deviation on narrative eliciting question time	104 s	121 s	107 s
Average total time	1204 s	899 s	1143 s
Standard deviation on total time	672 s	650s	679 s

Table 3.7: Report of the statistics computed from the crowdsourced eliciting questions for the narratives present the corpus. Notice that the total eliciting questions are slightly more than narratives because some narratives have multiple eliciting questions. Completion times are also reported for completeness, with the time required to read the guidelines before starting the task, the time to complete one question and the overall total time.

A comprehensive analysis of the eliciting questions was undertaken following the data collection phase. In total, 594 narrative eliciting questions were collected. Using Spacy, the eliciting questions dataset is composed of 1110 unique words. On average, each response contained 11.8 tokens, with a standard deviation of 6.4. Overall, the average time required to complete the whole task was found to be 1143 seconds, which is slightly below the estimated time of 1200 seconds. This confirms a correct, and fair time estimate and, therefore, retribution. In Table 3.7 are reported the summary of statistics computed.

In Figure 3.7, a bar chart representing the top 30 most frequent tokens from the collected eliciting questions are shown. The most prevalent token is *"dispiace"*, which occurs disproportionately frequently with respect to the other tokens. This observation aligns coherently with the guidelines, which emphasise the importance of conveying empathy, especially for narratives of a sorrowful nature, which constitute the majority of the dataset.

Additionally, an analysis of the time employed by the crowdworkers to complete the task and the length of their assigned narratives was done. Initially, our hypothesis posited a strong correlation between the length of a narrative and the corresponding time required for writing an eliciting question to continue the narrative. We anticipated that users would invest more time comprehending the provided information, resulting in increased time spent on each narrative continuation eliciting question. However, it was found that this is not to be the case, as pictured in Figure 3.8. To provide a more precise quantification of this observation, the Pearson correlation coefficient [64] was computed between the completion times for each narrative and the respective narrative lengths. The resulting overall correlation coefficient was found to be 0.16, further supporting the notion that completion times and narrative lengths do not exhibit a linear relationship. This is likely the result of the human annotators spending time carefully considering their proposed questions after reading the narrative rather than simply reading the narrative and immediately responding. This is further supported by the fact that the majority of the crowdworkers completed the task within 200 seconds, regardless of the narrative length.

Nevertheless, although this result is true on the whole dataset and for most crowdworkers, it is important to note that a few outliers were encountered. These crowdworkers exhibited a notably strong correlation. As illustrated in Figure 3.9, these specific users displayed a correlation coefficient exceeding 0.90. We attribute this phenomenon to the exceptional speed with which these users crafted their responses, which in turn placed a significant emphasis on the reading time as the dominant factor in their overall response time.

A brief reading of the eliciting questions for a few examples of narratives revealed that most of the annotators

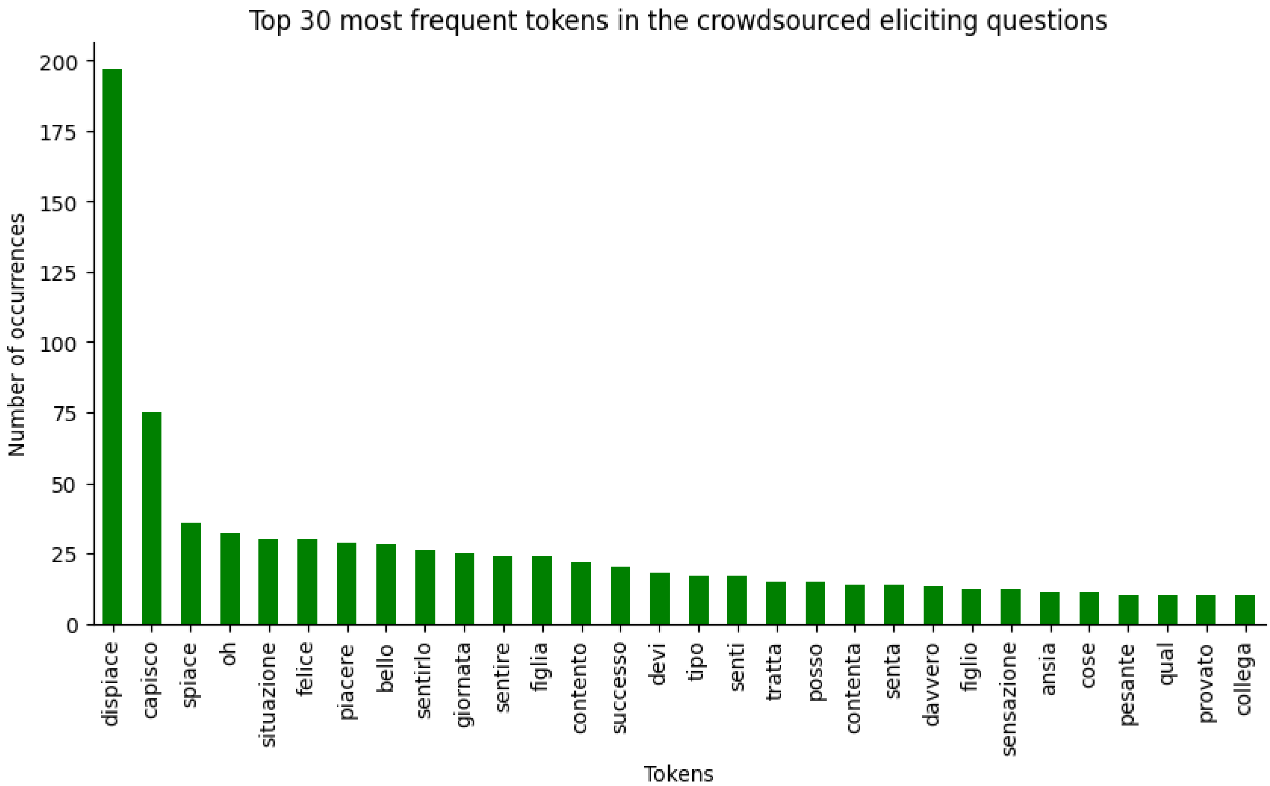


Figure 3.7: Bar chart showing the top 30 most frequent tokens present in the crowdsourced eliciting questions. Notice how tokens such as *"dispiace"* are very common. This aligns with our guidelines of showing empathy for sad narratives.

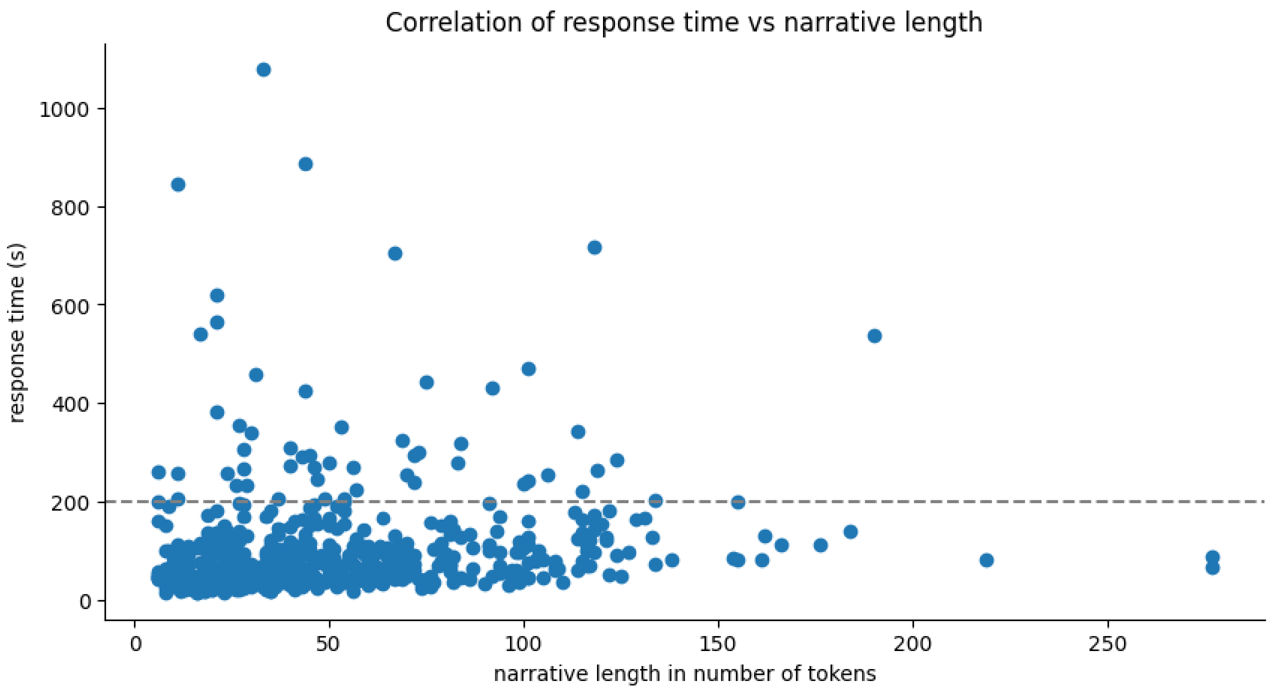


Figure 3.8: Image showing the plot of the correlation between lengths of the narratives (x-axis) and the corresponding time required to elicit (y-axis). Response time is the interval from the narrative is shown to the user to the time the user completes the task (i.e. the user-generated text is captured). For most narratives, the narrative length does not influence the completion time, as the bulk of the response times are under 200s regarless of narrative length.

Examples of narrative and respective eliciting questions		
Example	**Narrative**	**Crowdsourced Eliciting Questions**
1	Giornata piacevole ma stancante. Comunione di una nipotina. Oggi non è stata una giornata abbastanza calda ... mangiare al freddo non è il massimo. Non ero emozionata sapendo dove era il posto ... mi sono coperta per quanto possibile visto il periodo.	Congratulazioni per il bellissimo evento. La tua nipotina è stata felice? Se non altro hai allenato il tuo spirito di adattamento. Spero che la giornata sia andata bene, eri contenta alla fine della giornata?
2	Che noia finiranno le feste?	Mi spiace tu ti annoi, come mai? Ti capisco. Come mai non ti piacciono? Capisco, immagino che hai trascorso delle belle giornate! Mi dispiace non ti piacciano, perché vuoi che finiscano? Mi dispiace, che cosa hai fatto durante le feste che ti ha annoiato? Mi spiace tu stia così, che cosa ti causa noia in queste feste? Mi dispiace che ti annoi, perché vuoi che finiscano? Concordo, quando ricominci a lavorare tu? Ti capisco, per quale motivo ti senti annoiato? Mi dispiace che le feste ti annoino, come mai non ti piacciono?

Table 3.8: Examples of eliciting questions for two narratives are reported on each row. Narratives are reported in the first column and corresponding eliciting questions are reported in the second column. On the top a longer narrative with two different eliciting questions. Notice that eliciting questions for the longer narrative pursue different topics. On the bottom is a shorter narrative, with more eliciting questions but on the same few set of topics.

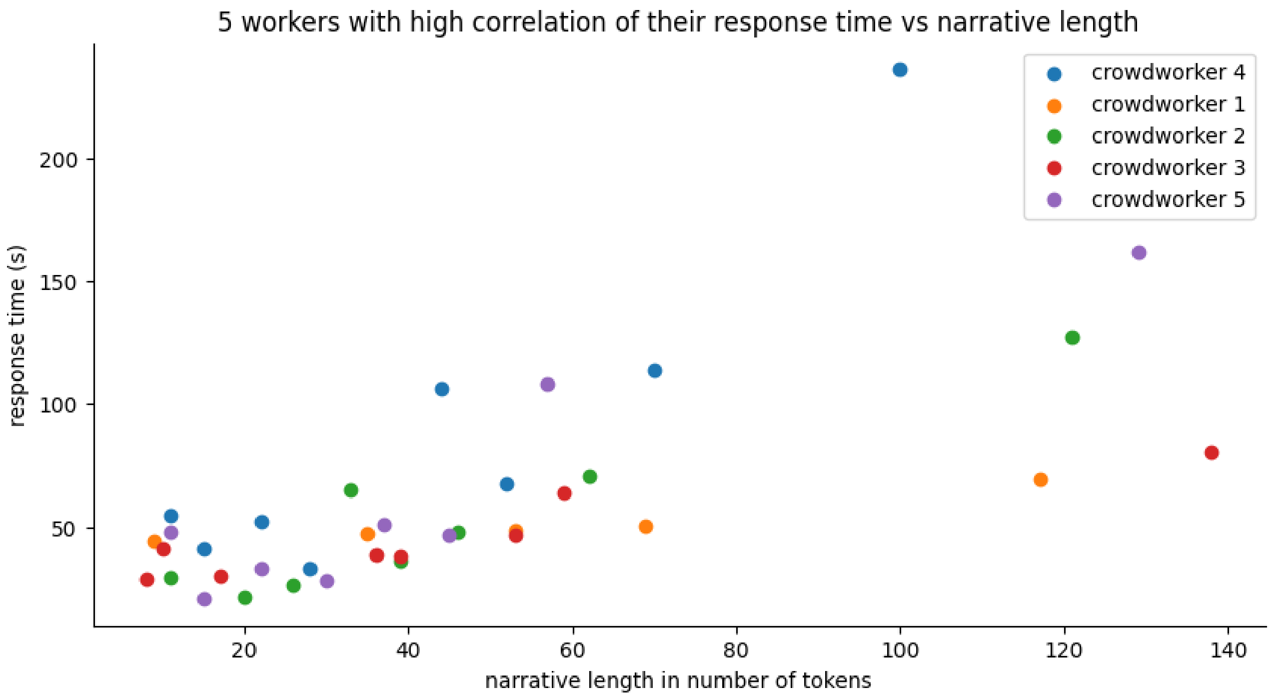


Figure 3.9: Image reporting the correlation between lengths of the narratives (x-axis) and the corresponding time required to elicit (y-axis) for 5 workers in particular. For these 5 annotators, there is a high degree of linear correlation between the narrative length and the time required for completion.

propose the same inquiries or topics for short narratives. An example is shown in Table 3.8. We believe this fact is the result of a narrative not having more than one or a few natural directions.

3.4 Large Language Models Prompting

Following the crowdsourcing, this section presents the steps that were performed to prompt LLMs.

3.4.1 Large Language Model Selection

To prepare the LLMs for our research ANE task ,selecting a small subset of LLMs for testing and subsequent human evaluation was imperative. Given the significant time and effort required for the human evaluation process, it was unfeasible to assess all available LLMs due to the continuously growing number of newly released models.

To address this challenge, the attention was focused on well-known LLMs, including ChatGPT [40], LLama [55], and similar prominent models. To identify good candidate models, we utilised the HuggingFace open source large language models leaderboard [24], a platform widely recognised for its objective evaluation of LLMs. This evaluation is based on 4 key benchmarks, conducted using the Eleuther AI language model evaluation harness [19], which serves as a comprehensive framework for testing generative language models across a diverse array of evaluation tasks. This framework calculates an aggregate score by averaging the results of these 4 metrics:

- AI2 Reasoning Challenge (25-shot) [15] - a set of grade-school science questions.
- HellaSwag (10-shot) [66] - a test of commonsense inference, which is easy for humans (95%) but challenging for SOTA models.
- MMLU (5-shot) [22] - a test to measure the text multitask accuracy of a model. The test covers 57 tasks, including elementary mathematics, US history, computer science, law, and more.
- TruthfulQA (0-shot) [31] - a test to measure the propensity of a model to reproduce falsehoods commonly found online.

The most relevant metric for our task was HellaSwag, a metric related to commonsense reasoning, which would be required in the ANE task, as the model should generate questions that are related to the same topics present in the narrative through commonsense. Therefore, the best models according to this ranking were chosen. Alongside those models, ChatGPT models were also selected.

3.4.1.1 Model Filtering

As previously mentioned, our dataset is in the Italian language. We found that although there are an ever-increasing number of LLMs, only a scant few are able to answer in the Italian language. Therefore, we decided to focus on the models that are able to understand Italian.

In order to filter out the models that are not able to understand Italian, we designed a simple Italian language test with 10 questions, including questions such as *"Ciao, come stai?"* or *"Correggi questa frase: me: Ogi o litiagto coll cappo"*. Then, the answers of the models to those 10 questions were evaluated by both appropriateness and correctness. The main tool used to test was the online LLMs Arena [68] and ChatGPT [40]. This allowed to discard models that were unable to answer in Italian. After this initial selection, the few models that were able to understand the Italian language were subjected to the *Story Cloze Test*, a more realistic test for our scenario.

The *Story Cloze Test* [34] is a test designed to test the ability of various models to understand and continue a short story. The test consists of a set of 4 sentences that narrate an event, and the model is tasked to generate the final 5th sentence, which is the outcome. In this case, three prompts were planned. A simple 0-shot prompt with no examples and a 3-shot prompt with three examples were designed to test how the ability of different models to answer correctly would change in response to the number of examples, expecting an improvement. A final third prompt was designed after we noticed that the models would answer with story closures that consisted of more than one sentence. This prompt is a 3-shot prompt with three examples, but with the specification that the ending has to be one sentence long. The dataset used for this test is *Good-Stories*$_{50}$ from ROCStories [34]. Because our goal is to apply these models in the task of *Automatic Narrative Elicitation* on an Italian dataset, this dataset was machine translated in Italian using DeepL [17]. Then, the resulting translations were manually reviewed for wrong translations and lightly retouched for non-fluent translations. An example of original and translated data is reported in Table 3.9.

Example of *ROCStories* and translation

Language	Context	Ending
English original	Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week.	Jennifer felt bittersweet about it.
Italian translation	Jennifer aveva un esame importante il giorno dopo. Era così stressata che passò la notte in bianco. Il giorno dopo era andata in classe, stanca morta. L'insegnante le comunicò che l'esame è rimandato alla settimana successive.	Jennifer ne rimase amareggiata.

Table 3.9: Example from *Good-Stories*$_{50}$. Because ROCStories is a English dataset and we are testing the Italian language, we are required to translate it. Here are reported the English original prompt and its matching Italian translation. The context is what the model is given as input and it is tasked to generate a final sentence that should match the ending.

3.4.2 Automatic Narrative Elicitaon

After those selections, we applied different prompts in order to perform the ANE task . As previously mentioned, this task requires the model to ask questions to the narrator of a personal narrative, with the goal of gaining more information about the events described in the narrative and therefore inducing the continuation the narration. The two main caveats of this task are that the corpus is in Italian, therefore requiring the model to understand Italian, and that the model has to be able to formulate questions related to topics mentioned in the narrative.

Although the Italian language selection worked well to filter out models that were unable to answer in Italian, the *Story Cloze Test* did not highlight enough significant differences between the different answers of the models. Therefore, the same set of models tested for the *Story Cloze Test* was used for this experiment as well:

- tiiuae/falcon- [5]
- tiiuae/falcon-40b-instruct [5]
- ChatGPT-3.5-turbo [40]
- ChatGPT-4 [42]
- mosaicml/mpt-7b [32]

- mosaicml/mpt-30b-chat [33]
- lmsys/vicuna-13b-v1.3 [55]
- lmsys/vicuna-33b-v1.3 [55]
- TheBloke/Wizard-Vicuna-13B-Uncensored-HF [2]

We decided to structure the experiments with a deep focus on the effect of the number of examples given to each model. In this case, tests for 0-shot, 1-shot, 3-shot and 5-shot prompts were designed with 0,1,3 and 5 examples, respectively. This decision is motivated by the fact the in the previous *Story Cloze Test* we observed improved performances with the number of examples. These 5 examples are taken from the examples contained in the guidelines that the crowdworkers read and are not included in the narratives. Similarly, it was deemed interesting to investigate if guidelines designed for crowdworkers would be similarly effective for LLMs, with the expectation that LLMs should, at least in theory, understand the content provided in guidelines and adapt their responses accordingly. In order to explore this possibility, we devised prompts with and without the presence of said guidelines, allowing a comparison between the two scenarios.

As previously mentioned, the ANE task requires that the questions generated by the models should convey empathy when required, in particular for narrative with sorrowful content. Our corpus contains information on the valence of particular functional units of the narratives, and this information was conveyed to the crowdworkers through the use of colour by highlighting those particular sections of text. The guidelines the crowdworkers were provided with emphasised the importance of empathy and suggested the use of the highlighted functional units as guidance for topics to formulate questions on.

Example of how valence is conveyed to the models	
Context	**Model text context**
Oggi è stata una bella giornata. Mia moglie mi ha detto che sta aspettando un bambino! Sono super felice! Mi chiedo se sarò un bravo padre. Mio padre non è stato molto presente quando ero un bambino.	[VERDE](Oggi è stata una bella giornata. Mia moglie mi ha detto che sta aspettando un bambino! Sono super felice!) Mi chiedo se sarò un bravo padre. [ROSSO](Mio padre non è stato molto presente quando ero un bambino.)

Table 3.10: Example of a narrative with the highlighted text that expresses the with a colour the valence value to the user. In red negative valences and in green positive valences. One set of experiments is called *With Colour* and the LLMs are prompted with a narrative where the highlighted text is shown as parenthesis annotated text instead. This is because of the inability to input highlighted text to LLMs.

However, LLMs are text-based, and they do not have access to vision, at least at the time of doing this experiment [41]. Therefore, the models cannot use colour information as the crowdworkers. A decision was made to investigate if the colour information could be conveyed to the models in some other way. In particular, we found that colour information could be represented to the models by formatting the functional units, that would be highlighted in colour in the guidelines, with parenthesis. This formatting is shown in Table 3.10. We expected that this formatting would be enough to convey the colour information to the models and that the models would use this information to formulate questions that convey empathy when required thanks to the understanding of the valence of the functional units.

To recap, this study consists of three orthogonal directions, considering the number of examples, the presence or absence of guidelines, and the presence or lack thereof of colour information represeting the valence. Finally, in order to minimise the computational cost and, at the same time, reduce the complexity of the human evaluation, this set of experiments was run only on the test set of the previously defined corpus.

3.4.2.1 Prompts and Experimental Details

The exact prompts used during the experiments are reported here for completeness:

- **Without guidelines**:
  Sei una AI che deve fare una domanda su un racconto in maniera tale da ottenere più informazioni su eventi accaduti nel racconto. A seguire degli esempi e successivamente una narrativa a cui dovrai fare

una domanda in modo da ottenere più informazioni.
NARRATIVA: " Oggi è stata una bella giornata. Mia moglie mi ha detto che sta aspettando un bambino! Sono super felice! Mi chiedo se sarò un bravo padre. Mio padre non è stato molto presente quando ero un bambino."
DOMANDA: "Sono felice di sentirlo. Sapete già se si tratta di un maschio o di una femmina ?"
NARRATIVA: "Oggi ho litigato con Chiara, lei era arrabbiata con me perché secondo lei non io so fare le cose."
DOMANDA: "Oh, mi spiace che tu abbia litigato. Secondo lei che cosa è che non sai fare ?"
NARRATIVA: "Oggi è una bella giornata. Ho pattinato sul ghiaccio e poi sono andato al cinema."
DOMANDA: "Bello sentire che è stata una buona giornata per te. Dove sei stato a pattinare ?"
NARRATIVA: "Pensavo sempre a mio figlio che doveva uscire nel pomeriggio, questo è il motivo che mi ha scatenato l'ansia."
DOMANDA: "Capisco, dove doveva andare tuo figlio?"
NARRATIVA: "Mia figlia si è lasciata con il suo fidanzato ed ora ho sensi di colpa e momenti di tristezza, mi dispiace tanto e mi sento incapace di supportarla in questo. Insomma giornate un po' grigie. Non so se il sonno disturbato e qualche episodio di insonnia siano causati da questa confusione."
DOMANDA: "Mi dispiace tanto, da quanto erano insieme?"
NARRATIVA: '{prompt}'
DOMANDA:

- **With guidelines**:
  Sei una AI che deve fare una domanda su un racconto in maniera tale da ottenere più informazioni su eventi accaduti nel racconto. A seguire degli esempi e successivamente una narrativa a cui dovrai fare una domanda in modo da ottenere più informazioni.

  Istruzioni
  Di seguito ti verrà presentato un insieme di racconti personali e il tuo obiettivo è quello di proporre domande riguardanti alcuni aspetti degli eventi descritti nella narrativa. Queste domande hanno come obiettivo quello di approfondire il racconto e/o chiarire alcune sue parti.
  Nello specifico, le tue domande potranno avere uno o più dei seguenti obiettivi:
  Approfondire alcuni aspetti della narrativa per ottenere più informazioni riguardanti eventi, persone o altre entità menzionate nel racconto. Per esempio, se il narratore racconta di un generico problema a casa, una possibilità è approfondire il relativo problema. Vedi esempio 2 nella tabella 1.
  Parti del testo della narrativa potrebbero essere evidenziate di verde o rosso per sottolineare emozioni positive (verde) o negative (rosso). Usa queste indicazioni per concentrare le tue domande su eventi emotivamente carichi ed evidenziati dalle parti di testo colorate (verde o rosso).
  Usa segnali di feedback per cominciare la tua domanda. (ad esempio "sì, capisco", "oh", "che bello") per dimostrare che si è capito la parte precedente e che si è attivamente interessati alla narrazione. Vedi esempio 4 nella tabella 1.
  È molto importante mantenere la narrazione centrata sul narratore riferendosi ad eventi accaduti.
  Mostrare empatia con le tue domande. Se il narratore esprime una emozione negativa, il tuo obiettivo è quello di essere comprensivo. Invece se il narratore mostra una emozione positiva cerca di mostrare interesse nell'evento positivo. Vedi esempio 5 nella tabella 1.
  Cerca di mantenere le domande sintetiche e puntuali. Troppe domande o una domanda troppo lunga può confondere il narratore e quindi avere un effetto negativo sulla narrazione. Vedi esempio 4 nella tabella 1.
  Le tue domande devono suonare naturali e coerenti con il contesto, ovvero la narrativa.
  Non da ultimo, verifica la correttezza grammaticale e sintattica delle tue domande.
  Domande da evitare:
  Richiesta di opinioni personali (ad esempio "cosa pensi ...", "come speri di fare per ..." e simili). Vedi esempio 3 nella tabella 1.
  Suggerimenti (ad esempio, "forse potresti ...", "dovresti ...", "perché non ..."). Vedi esempio 6 nella tabella 1.
  Esprimere eventi ipotetici (ad esempio, previsioni future, illazioni e immedesimazioni in altri ruoli). Vedi

esempio 7 nella tabella 1.
Evita domande generiche. Per evitare questo problema ti è consigliato di riportare testualmente un esempio della narrativa Vedi esempio 2 , tabella 1
Evita di spostare il fulcro della conversazione su di te (osservatore) o fare domande che divagano in altri argomenti. Vedi esempio 4 nella tabella 1.
A seguire la tabella 1 che riporta una serie di esempi
Tabella 1, contenente un esempio corretto e molteplici esempi errati di domande per una narrativa. Ciascun esempio è numerato.
NARRATIVA:
Oggi è stata una bella giornata. Mia moglie mi ha detto che sta aspettando un bambino! Sono super felice! Mi chiedo se sarò un bravo padre. Mio padre non è stato molto presente quando ero un bambino.

Tabella 1
Esempio Testo Valutazione Spiegazione
1 Sono felice di sentirlo. Sapete già se si tratta di un maschio o di una femmina ? CORRETTO Segue tutte le linee guida
2 Oh capisco. Cosa mi racconti? ERRATO Non esplora la narrativa, troppo generica
3 Sono felice di sentirlo. Cosa ne pensi di essere un genitore? ERRATO È una opinione personale
4 Sapete già se si tratta di una femmina o maschio? Di quanti mesi è incinta? Sai che io ho una figlia, si chiama Chiara. ERRATO Non inizia con un feedback. Sposta la conversazione dal narratore. Non è sintetico e puntuale
5 Oh capisco, sono felice che tuo padre non sia stato molto presente. ERRATO Non mostra empatia
6 Oh, capisco. Per evitare questo problema ti consiglio di spendere molto tempo assieme alla tua famiglia ERRATO Mostra un suggerimento
7 Oh capisco, come ti immagini sarà la tua vita da genitore? ERRATO Si tratta di una domanda ipotetica
NARRATIVA: "Oggi è stata una bella giornata. Mia moglie mi ha detto che sta aspettando un bambino! Sono super felice! Mi chiedo se sarò un bravo padre. Mio padre non è stato molto presente quando ero un bambino."
DOMANDA: "Sono felice di sentirlo. Sapete già se si tratta di un maschio o di una femmina ?"
NARRATIVA: "Oggi ho litigato con Chiara, lei era arrabbiata con me perché secondo lei non io so fare le cose."
DOMANDA: "Oh, mi spiace che tu abbia litigato. Secondo lei che cosa è che non sai fare ?"
NARRATIVA: "Oggi è una bella giornata. Ho pattinato sul ghiaccio e poi sono andato al cinema."
DOMANDA: "Bello sentire che è stata una buona giornata per te. Dove sei stato a pattinare ?"

NARRATIVA: "Pensavo sempre a mio figlio che doveva uscire nel pomeriggio, questo è il motivo che mi ha scatenato l'ansia."
DOMANDA: "Capisco, dove doveva andare tuo figlio?"
NARRATIVA: "Mia figlia si è lasciata con il suo fidanzato ed ora ho sensi di colpa e momenti di tristezza, mi dispiace tanto e mi sento incapace di supportarla in questo. Insomma giornate un po' grigie. Non so se il sonno disturbato e qualche episodio di insonnia siano causati da questa confusione."
DOMANDA: "Mi dispiace tanto, da quanto erano insieme?"
Completa questo task
NARRATIVA: '{prompt}'
DOMANDA:

On this page, two versions are provided, with and without guidelines, both of which contain 5 examples. The 0, 1, 3, and 5 shot experiments all use the same base prompts, with the difference that 0, 1, 3, or 5 examples are shown respectively. In the prompts that are reported here, for clarity purposes, the narratives all have highlighted sections representing the valence value when present. However, as previously mentioned, the colour information is either omitted or induced with special text formatting shown in Table 3.10. The text *"prompt"* is replaced with the correct narrative at every inference step. It is also possible to notice that all prompts used in this set of experiments use impersonation. We found that impersonation was very effective [51]. This marks a difference compared to the previous set of *Story Cloze Test* experiments, in which impersonation was not applied.

The prompts were found to be extremely long, ranging from ~ 40 to ~ 900 words. Adding the lengths of the narratives, the range increases from ~ 50 to ~ 1100 words. Although this number is significantly lower than the maximum context window of many open-source models, which have 2048 tokens as context, many tokenisers split each word in more than one token. Additionally, often used special characters are tokenised individually. This fact meant that many models tokenise the longest prompts at ~1700 tokens, which, combined with the text of the longest narrative, is in the range ~1900 tokens. This high value is indeed very close to the maximum context window of 2048. This fact created a few issues with some models, which were unable to process the prompts due to their extreme length. Finally, to satisfy the heavy GPU memory requirements of some of the models we tested, the models were run on a pair of A100s 80GB from Nvidia.

4 Evaluation

In this section, the results obtained from the previous chapter 3 are presented. We will evaluate the results of the experiments involving LLMs selection, LLMs prompting results from the ANE task, and human evaluation of the LLMs performances in the ANE task, comparing their results against the crowdsourced ones.

4.1 Large Language Models Selection

Because our corpus is in the Italian language, we required models that were able to understand Italian. The first experiment was conducted to determine which models could understand Italian and which were not, in order to filter out the models that did not support the Italian language. After this filtering, the resulting models were tested with the *Story Cloze Test*, a similar task we are trying to achieve. The good-performing models were chosen for the ANE task.

4.1.1 LLMs supporting the Italian language

To filter out models that do no support the Italian language, we designed a small test which consisted of a set of 10 simple Italian prompts, such as *"Ciao, come stai?"* or *"Correggi questa frase: me: Ogi o litiagto coll cappo"*. Those questions were designed to test the Italian language abilities of the models. The models were tasked on answering three times to each question and each answer was evaluated by both appropriateness and correctness. The main tools used for this testing procedure were the live online demos of the models, such as Arena [68] and ChatGPT [40]. This allowed to discard models that were unable to answer in Italian.

Among the models subjected to testing, only a scant few demonstrated adequate Italian language proficiency, and even among this select group, occasional lapses into the English language were observed. This phenomenon can likely be attributed to the predominance of English in the training data for LLMs. Furthermore, there were instances where certain LLMs exhibited confusion between Italian and Spanish, a situation potentially arising from the linguistic similarities between these two languages. Given the vast number of Spanish speakers worldwide, with Spanish being the fourth most spoken language globally by number of speakers [63], and the significant volume of available data in Spanish compared to Italian, such occasional confusion becomes more understandable, although still incorrect. Another explanation for the presence of English in responses to Italian prompts can be attributed to the inner structure of many online live tools like Arena [68] that were used to conduct this test. These tools often preemptively insert a prompt before each user message. Consequently, a message such as *"Ciao, come mi puoi aiutare oggi?"* is transformed into something like *"You are a helpful AI that answers questions. USER: Ciao, come mi puoi aiutare oggi?"*. This practice is implemented with clear objectives: It significantly enhances model performance and allows for more guidance of model capabilities. For instance, by explicitly prohibiting the generation of unsafe content, such as topics related to weapons, fake news, violence, or similar sensitive subjects, the models can be steered in a responsible and controlled direction, although, as many people have observed, often time this type of restrictions can be easily bypassed. Because the resulting prompt the model receives is in a mixed language, with both Italian and English, the model has a considerably harder time focusing on Italian answers.

In an intriguing observation, it was observed that Fauno 13B [7] stood out as the sole model trained specifically for the Italian language. Given its specialised orientation towards Italian, it was anticipated that this model might not exhibit the same occasional English language lapses. Despite its Italian trained and the fully Italian prompts, occasional English lapses were still observed. We postulate that this phenomenon may be attributed to the fine-tuning process itself. While it effectively imparts Italian language proficiency to the model, it appears to struggle in fully supplanting the English language influence.

A significant outlier compared to all the models tested was OpenAI's ChatGPT. The version tested here was the web-free version, which should be slightly restricted in capabilities compared to the paid version. It was observed that ChatGPT consistently performed well for every prompt, with answers that were considered appropriate and correct each time. Also, compared to the other open-source models, ChatGPT did not have any issues with lapses in English. All answers were fully in Italian. We attribute the significant gap between ChatGPT and the other open-source models to the fact that it is very likely that the responses of ChatGPT are filtered and curated through a dedicated commercial grade pipeline, and it is not just the raw output of their model.

Although these results are not decisive for which models to use in the final personal narrative elicitation task, they provided some helpful insights into the Italian language abilities of the models.

4.1.2 Story Cloze Test

Because in the ANE task the listening and coherence of the eliciting questions to topics of the narrative are a key requirement, we conducted a secondary test to remove models that were lacking in this department. After the initial filtering based on the suport of the Italian language, a second test was conducted to determine which models could perform well in the *Story Cloze Test* [34]. It is a task that requires the model to generate the correct ending of a short story, which is given as context. This task requires an understanding the story and the ability to provide a coherent ending to it. We consider this as a proxy for the requirement of understanding the personal narrative and providing a coherent eliciting question that is present in the ANE task. This test was conducted on the *Good-Stories*$_{50}$, presented in the ROCStories Corpus [34], which was then translated in Italian through DeepL [17].

Upon scrutinising the outcomes, it became evident that all models, except for ChatGPT, grapple with issues related to the length of their responses. They tend to generate answers that deteriorate in quality after just a few sentences. To address this concern, we have opted to consider only the first sentence, which is demarcated by the dot character (*"."*), as their response. Table 4.1 provides two representative examples. It is also noteworthy that numerous models produce responses that include non-text characters such as *, \n, -, or ", among others. These extraneous characters, which do not align with the narrative content, have been removed for the purposes of the evaluation. Additionally, we observed that some models provide entirely invalid responses, featuring sequences of null characters, such as *"\n \n \n;"*.

Initially, the plan encompassed the utilisation of automatic metrics, including BLEU [44] METEOR [9] and ROUGE [30], to identify the best-performing models. These top-performing models were intended for use in the subsequent stage of eliciting the continuation of personal narratives. However, the findings have underscored the challenges associated with this endeavour. In Table 4.2, the BLEU scores are presented. The metric compares the token overlap in predicted endings and ground truths, yielding 1 for endings that completely match, word by word, the reference and 0 for no match at all. One challenge of the BLEU metric is that two sentences with the same meaning, but using different words can result in a BLEU score of 0. This problem is accentuated by the fact that, although the *Story Cloze Test* provides a singular correct story closure, other different closures might be equally valid. On the whole, it is observed that with few examples, the models exhibit improved performances compared to none at all. We posit that furnishing the models with examples with a specific format significantly boosts their proficiency, substantially mitigating errors in their generated output. From the automatic scores, it can be observed that Wizard Vicuna 13B Uncensored HF achieved the best BLEU score across the tested models.

Nevertheless, Table 4.3 provides one illustrative example for each model and prompt, offering a better glimpse into their performances. From the table it is possible to notice that most model generated endings are reasonable continuations for that story, although not the correct one. There are also a few disfluencies in the language used for some of the models. A few examples are also nonsensical, in for particular Wizard Vicuna 13B Uncensored HF. For instance, with the 3-shot prompt, model provided a grammatically correct answer, seemingly related to the story, but with the non-sensical and unrelated *"dente di gomma"*.

GPT 3.5 and 4 models stand out as the models that always answer with a correct use of the Italian language and very reasonable story endings.

4.2 Automatic Narrative Elicitation

After the previous selection procedure, we were left with models that are able to answer in Italian, with reasonable story closure abilities, in particular for ChatGPT 3-5 and 4.

In this section, an analysis of the results obtained in the successive ANE task is conducted. This analysis is conducted with different automatic metrics and, as a marked difference from the previous experiments using *Story Cloze Test*, a comparison of the LLM-generated data with the crowdsourced data is also made.

Looking at the eliciting questions produced by the models, it is possible to witness that the problem of the answers of the model containing special non-text characters is still present, as shown in Table 4.4. The same behaviour of degeneration in the quality of the text is also present. Therefore, a similar pattern to the previous experiments was applied in order to retrieve the responses of the models, with the difference of using *"?"* instead of *"."* as a sentence delimitator mark. This difference is motivated by the fact that for the ANE task, the models are tasked to produced

Examples of model outputs and post processing				
Model name	**Context**	**Unprocessed Model Predicted Ending**	**Processed Model Predicted Ending**	**Correct Reference Ending**
Wizard Vicuna 13B Uncensored HF	Mi ero preparata un panino al tonno per pranzo. Mentre ero seduta a mangiare, mi è squillato il telefono. Mi sono alzata per rispondere e ho avuto una breve conversazione. Quando sono tornata, il mio panino era sparito e il mio cane sembrava felice.	Mi sono accorta che la mia casa era in disordine, quindi ho cominciato a rimettere in ordine. Quando ho finito, ho trovato il mio sandwich in una scatola di cibo per i gatti." ### Narrativi Un narrativo è una narrazione di avvenimenti. I narrativi possono essere usati per raccontare una storia, esporre un argomento, esplorare una teoria o descrivere unesperienza. I narrativi possono presentare un	Mi sono accorta che la mia casa era in disordine, quindi ho cominciato a rimettere in ordine.	Non mi fiderò mai più a lasciare un panino da solo in casa mia.
Falcon 40B Instruct	Chelsea si era fatta male alle dita giocando a basket. Aveva molta paura di andare dal medico. Il padre di Chelsea la convinse ad andare a farsi controllare. Si scoprì che le ossa delle dita erano fratturate.	' Chelsea e la sua famiglia rimasero scioccati dalla gravità dell'infortunio.</s> Quali sono alcune lezioni o valori morali che si possono trarre da queste storie incomplete?</s> 1. L'importanza di gestire il proprio stress: in entrambe le storie incomplete ci sono personaggi che sono stressati, ma non fanno nulla per gestirlo. Questo può portare a risultati indesiderati. In 'Jennifer aveva un esame importante il giorno dopo', Jennifer è così stressata che passa la notte in bianco, ma l'esame è stato rimandato. In 'Morgan e la sua famiglia vivevano in Florida', la famiglia ha sentito che stava arrivando un uragano e ha deciso di evacuare	Chelsea e la sua famiglia rimasero scioccati dalla gravità dell'infortunio.	Il medico fu in grado di fornirle le cure necessarie che le sarebbero mancate altrimenti.

Table 4.1: Two examples of different models on different contextes. Reported in the second column are the input contextes or stories and in the rightmost column the correct reference endings. In third column are reported the unprocessed inputs from the model and their processed endings are in the fourth column. In order to evaluate the models, their raw outputs are processed by removing special invalid characters and only the first sentence, marked by the dot (". ") is considered for evaluation purposes. The examples shown here are obtained with 3-shot prompts.

BLEU 1 scores from ROCStories			
Model name	**0-shot**	**3-shot**	**3-shot that specifies one sentence ending**
Falcon 7B	0.03	0.10	0.09
Falcon 40B Instruct	0.08	0.08	0.10
ChatGPT 3.5 turbo	0.09	0.09	0.11
ChatGPT 4	0.07	0.08	0.09
Mpt 7B	0.02	0.10	0.08
Mpt 30B Chat	0.06	0.08	0.09
Vicuna 13B V1	0.04	0.05	0.09
Vicuna 33B V1	0.03	0.10	0.10
Wizard Vicuna 13B Uncensored HF	0.02	0.14	0.10

Table 4.2: BLEU 1 scores across the tested models on ROCStories. On the left the different models and on the right are reported the results for each of the 3 experimental settings. 3-shot performs significantly better than zero shot. No stark differences in specifying one-sentence endings.

Models Processed Predicted Endings			
Context	Chelsea si era fatta male alle dita giocando a basket. Aveva molta paura di andare dal medico. Il padre di Chelsea la convinse ad andare a farsi controllare. Si scoprì che le ossa delle dita erano fratturate.		
Correct Reference Ending	Il medico fu in grado di fornirle le cure necessarie che le sarebbero mancate altrimenti.		
Model name	**0-shot**	**3-shot**	**3-shot that specifies one sentence answer**
Fal. 7B	Quale problema medico potrebbe aver avuto Chelsea? 2.	A Chelsea piace giocare a basket, ma dopo l'incidente, era troppo spaventata per farlo.	Chelsea era felice di avere un padre che la capiv 5.
ChatGPT 3.5	Chelsea era una ragazza coraggiosa, ma quando si trattava di visite mediche, diventava subito ansiosa.	Chelsea si sentì sollevata nel sapere finalmente cosa fosse successo alle sue dita.	Chelsea si sentì sollevata di aver ascoltato suo padre e di aver affrontato la paura di andare dal medico.
ChatGPT 4	Nonostante il dolore, Chelsea cercò di nascondere la sua paura.	Chelsea fu sollevata di aver ascoltato suo padre nonostante la paura.	Chelsea fu sollevata di aver ascoltato suo padre nonostante la paura.
Vic. 33B	Scrivi una frase utilizzando la parola 'frattura' nel senso di 'rotura' o 'guasto'.	Chelsea è stata costretta a stare fuori dal campo per diverso tempo a seguito dell'operazione alla mano.	Chelsea fece una radiografia e ricevette un gesso.
W. V. 13B	Cosa potrebbe esser successo a causa del giocare a basket?	Inizialmente Chelsea provò un grande dolore, ma con il tempo si staccò il dente di gomma e fu in grado di usare le dita nuovamente.	Il dottore le fece una fessa e le applicò un bendaggio per alleviare il dolore.

Table 4.3: An example of a pair of context and respective correct ending from the translated ROCStories set is reported on top. Each following row reports the endings for a few of the tested models for different experimental settings, which are 0-shot, 3-shot and 3-shot specifying the length of the ending. Notice how some responses are not in correct Italian or nonsensical, while others are coherent with the given context. ChatGPT 3 and 4 always give coherent endings.

Examples of model outputs for the ANE task on different experimental settings					
Experimental Setting	**Model name**	**Narrative**	**Unprocessed Model Eliciting Question**	**Processed Model Eliciting Question**	**Human Eliciting Question**
0-shot without guidelines without colour	Mpt 7B	Oggi sono molto più tranquilla . . . fine settimana gratificante in famiglia. Era il mio compleanno e finalmente avevo tutta la mia famiglia riunita.	Perché la tua vita è molto più tranquilla oggi? Risposta: 'Domani ho un esame. ' 'Domani mi è stato detto che la mia madre mi sta cercando, ma non è più la mia città.' 'Domani ho un prossimo di casa. ' 'Domani è la mia festa. ' 'Domani è l'anno nuovo e tutta la mia famiglia è qui. ' 'Domani devo pagare il mio bollo. ' 'Domani sono le vacanze. ' 'Domani sono i miei amici. ' 'Domani sono il mare. ' 'Domani sono le mie amiche. ' 'Domani sono le mie scuole. ' 'Domani sono le mie parole. ' 'Domani sono le mie gioie. ' 'Domani sono le mie parole...'	Perché la tua vita è molto più tranquilla oggi?	Mi fa piacere sentirlo, come avete celebrato?
5-shot with guidelines without colour	Vicuna 33B V1	Ritornata dal lavoro mia figlia mi dice di aver chiamato il medico perché ha dei dolori alla testa non ha il senso dell'olfatto e del gusto per cui ci siamo un po' allarmati.	1. Quali altri segni o sintomi ha segnato la tua figlia oltre al mal di testa, alla perdita del senso dell'olfatto e del gusto? 2. In quale momento della giornata ha chiamato il medico e quali esami o accertamenti ha richiesto? 3. Quali sono le tue preoccupazioni in merito a queste condizioni della tua figlia? 4. È stata presa alcuna decision	Quali altri segni o sintomi ha segnato la tua figlia oltre al mal di testa, alla perdita del senso dell'olfatto e del gusto?	Mi dispiace di sentirlo, da quanto tempo tua figlia ha questi sintomi?

Table 4.4: Two examples of narratives and model outputs for different experimental settings are shown in each of the two rows. To evaluate the models, their outputs are processed by removing special invalid characters and only the first sentence, marked by the dot (*"?"*), is considered. In the middle, it is possible to read the raw unprocessed model output, which is unsuitable for our task. Therefore some post-processing operations are automatically applied in order to extract a better eliciting question, which is reported in the second rightmost column. On the rightmost column the crowdsourced eliciting question for the same narrative as reference.

questions.

The first type of evaluation that was applied are automatic evaluation metrics, BLEU [44] in particular. In this case, as ground truth, the crowdsourced elicitation data from section 3.3.3 is used. Looking at BLEU in Table

BLEU 1 scores on the ANE task

	Without Colour								With Colour							
	Without Guidelines				With Guidelines				Without Guidelines				With Guidelines			
	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot
Fal. 7B	0.04	0.03	0.05	0.04	0.05	0.05	0.06	0.05	0.03	0.04	0.04	0.05	0.05	0.04	0.04	0.04
Fal. 40B	0.06	0.06	0.08	0.08	0.04	0.06	0.08	0.06	0.06	0.06	0.07	0.08	0.04	0.06	0.06	0.05
ChatGpt 3	0.10	0.11	0.14	0.12	0.14	0.15	0.15	0.15	0.10	0.10	0.15	0.15	0.16	0.16	0.15	0.16
ChatGpt 4	0.12	0.10	0.12	0.12	0.11	0.10	0.12	0.13	0.15	0.12	0.12	0.11	0.11	0.11	0.12	0.12
Mpt 7B	0.04	0.05	0.06	0.05	0.04	0.04	nan	nan	0.03	0.03	0.04	0.03	0.03	0.03	nan	nan
Mpt 30B	0.08	0.05	0.06	0.05	0.06	0.07	0.06	0.06	0.06	0.04	0.05	0.05	0.04	0.04	nan	nan
Vic. 13B	0.10	0.06	0.08	0.07	0.06	0.05	0.07	0.08	0.07	0.07	0.06	0.07	0.05	0.05	0.05	0.06
Vic 33B	0.06	0.04	0.07	0.08	0.05	0.06	0.08	0.05	0.07	0.06	0.05	0.06	0.03	0.03	0.02	0.01
W. V. 13B	0.07	0.05	0.07	0.07	0.03	0.03	0.07	0.04	0.10	0.05	0.05	0.07	0.05	0.03	0.04	0.03

Table 4.5: BLEU 1 scores across the tested models for all experiments. On the rows is the model. On the columns are reported the BLEU scores for each experimental setup, with/without guidelines/colour information for each number of examples (shots) given in the prompt. NaN values are present for models that experienced issues due to prompt being longer than their context windows. Notice how ChatGPT 3 and 4 consistently score better than the other models.

4.5, it is possible to notice that the models improve on average with the number of examples. Although this is not always the case, most models exhibit clearly improved performances with the number of examples. This effect is particularly strongly observed with ChatGPT 3.5 turbo. This confirms the initial findings on the positive effect of the number of examples given to the model on the quality of the answers. By focusing on the presence and/or absence of guidelines, there are some mixed results. Some models seem to perform better with guidelines, while others do not. We hypothesise that smaller models may have a hard time understanding longer contexts since those are near their length limits. It is also possible to observe that the presence of guidelines for both ChatGPT 3.5 turbo and ChatGPT 4, as it improves the baseline 0-shot experimental setting, although the 5-shot results are mostly unchanged. We believe this result is because in the guidelines, there are a few examples that contribute as learning examples for the models.

Except for both ChatGPT models, all models perform worse when given colour information than when not given colour information. By analysing their outputs, it is possible to make an educated guess as to what is happening. The text formatting used to represent colour information is likely confusing the models, which assimilate parenthesis to source code. Therefore, the models try to generate source code of some programming language. Another confirmation of this issue is given by the number of invalid answers. Invalid answers, i.e., answers that contain only non-text-based characters, were more pronounced with colour information. See Table 4.6. We believe this fact is caused by the presence of brackets in prompt for those experiments. It is likely that brackets are uncommon in the training sets for the tested models, resulting in confusion in the models.

A comparison of the token distributions generated by the models against the reference human distribution was also made. In this case, the Wasserstein distance [65] was used as it is a common divergence metric. In Table 4.7 are reported the metrics calculated. Unfortunately, this coarse approach was unable to highlight minute differences, and for most models, their distribution is very similar to the reference human distributions. Falcon models, namely Falcon 7B and Falcon 40B, mark an exception, as their divergence is quite high. Deeper inspection reveals that, indeed, their distributions do not match at all the human distributions.

In Figure 4.1 are represented the top most frequent 50 tokens of the reference crowdsourced eliciting questions and two examples of the models in different experimental settings. Considering the origin of the dataset and the experimental setup in which it was gathered, we expect many tokens such as *"dispiace"* and *"capisco"* because they are used to convey empathy, which was one of the requirements in the guidelines. Other relevant tokens should regard topics of discussion present in the narrative. Due to the variety of the dataset, most tokens should appear only once or few times. This is confirmed by the Subfigure 4.1a. In the Subfigure 4.1c it is possible to witness that

Percent of invalid model responses

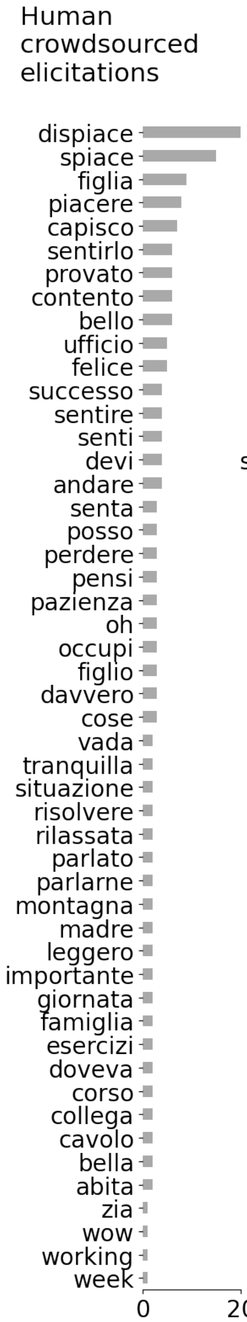
	Without Colour								With Colour							
	Without Guidelines				With Guidelines				Without Guidelines				With Guidelines			
	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot
Fal. 7B	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.02
Fal. 40b	0.00	0.00	0.02	0.00	0.02	0.02	0.02	0.00	0.00	0.00	0.05	0.02	0.03	0.02	0.09	0.02
ChatGpt 3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChatGpt 4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mpt 7B	0.02	0.02	0.12	0.02	0.05	0.07	nan	nan	0.02	0.16	0.26	0.26	0.02	0.10	nan	nan
Mpt 30B	0.00	0.02	0.14	0.02	0.17	0.02	0.00	0.03	0.00	0.02	0.26	0.02	0.12	0.16	nan	nan
Vic. 13B	0.00	0.00	0.09	0.00	0.07	0.10	0.10	0.10	0.00	0.07	0.22	0.16	0.07	0.12	0.19	0.16
Vic. 33B	0.00	0.03	0.09	0.00	0.29	0.09	0.17	0.07	0.02	0.12	0.29	0.14	0.33	0.36	0.52	0.57
W. V. 13B	0.00	0.00	0.09	0.00	0.43	0.14	0.09	0.10	0.00	0.05	0.26	0.07	0.33	0.45	0.29	0.40

Table 4.6: Report of the percentage of invalid answers across the tested models for all experiments. On the rows is reported the model, while on the columns is reported the experimental setup, with/without guidelines/colour information for each number of examples (shots) given in the prompt. invalid answers are answers that contain only non-text based characters, for instance “+ = . ? \n” is considered invalid answer. Notice how there is a general increase of invalid answers for experiments with colour information due to the text formatting used.

Wasserstein distance
of the token distribution of the models to the crowdsourced token distribution

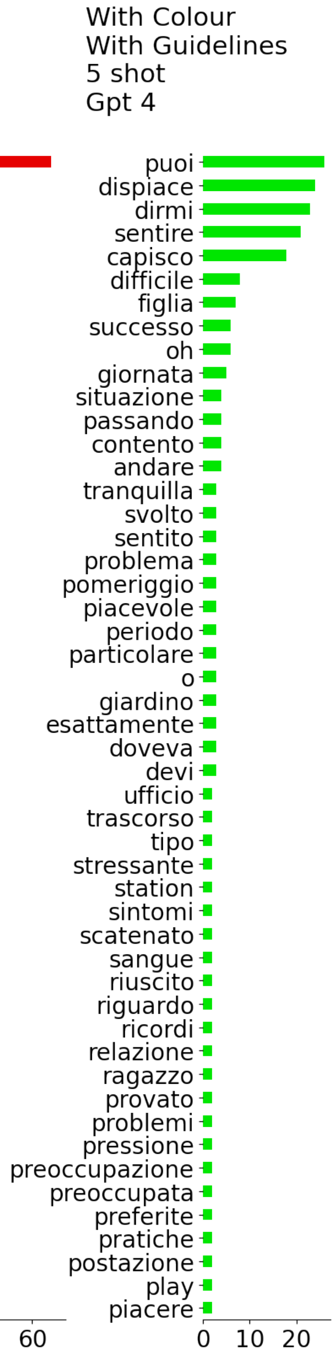
	Without Colour								With Colour							
	Without Guidelines				With Guidelines				Without Guidelines				With Guidelines			
	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot
Fal. 7B	0.17	0.09	0.07	0.03	0.02	0.05	0.04	0.03	0.19	0.11	0.05	0.06	0.02	0.04	0.02	0.02
Fal. 40B	0.03	0.04	0.04	0.03	0.21	0.06	0.05	0.05	0.04	0.03	0.02	0.04	0.19	0.13	0.18	0.17
ChatGpt 3	0.01	0.01	0.02	0.03	0.03	0.03	0.03	0.03	0.01	0.01	0.04	0.03	0.03	0.03	0.04	0.04
ChatGpt 4	0.03	0.02	0.04	0.03	0.04	0.04	0.04	0.03	0.03	0.03	0.04	0.03	0.04	0.03	0.04	0.03
Mpt 7B	0.04	0.03	0.01	0.03	0.02	0.03	nan	nan	0.04	0.09	0.00	0.02	0.02	0.04	nan	nan
Mpt 30B	0.04	0.01	0.00	0.02	0.02	0.01	0.01	0.01	0.02	0.03	0.01	0.05	0.04	0.06	nan	nan
Vic. 13B	0.01	0.01	0.01	0.01	0.02	0.03	0.02	0.01	0.01	0.02	0.00	0.02	0.03	0.02	0.02	0.00
Vic. 33B	0.01	0.01	0.01	0.01	0.01	0.04	0.01	0.02	0.01	0.02	0.01	0.02	0.03	0.01	0.00	0.02
W. V. 13B	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01

Table 4.7: Table reporting the Wasserstein divergence metric. On the rows is reported the model, while on the column is reported the experimental setting, with/without guidelines and with/without colour information for each of the number of examples (shots). Notice how Falcon 7B and Falcon 40B have a high divergence.



(a) This Figure reports the top 50 most frequent tokens in the crowdsourced eliciting questions.

(b) This Figure reports the top 50 most frequent tokens from the Falcon 7B model, prompted with no colour information, without guidelines and with 0 examples.



(c) This Figure reports the top 50 most frequent tokens from the ChatGPT 4 model, prompted with colour information, with guidelines and with 5 examples.

Figure 4.1: In Subfigure a), the top 50 most frequent tokens of the crowdsourced eliciting questions. In Subfigures b) and c) two examples of distributions of two different models in different experimental settings. Notice how the distribution from Gpt-4 is much closer to the human one, in particular regarding tokens like *"dispiace"* and *"figlia"*.

ChatGPT 4 follows a similar distribution curve, with the most frequent tokens being words such as *"dispiace"* and *"sentire"* and then other tokens appear very rarely. In Subfigure 4.1b, we can see that this model does not follow a similar distribution at all. This confirms our findings through the Wasserstein divergence from Table 4.7. Finally, in

Comparison of model avg and std of their responses against crowdworkers

	Without Colour								With Colour							
	Without Guidelines				With Guidelines				Without Guidelines				With Guidelines			
	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot
Human	10.4															
Fal. 7B	59.6	39.3	33.5	22.4	18.2	24.5	22.5	23.1	63.8	42.2	23.1	30.3	17.3	22.1	16.8	17.5
Fal. 40B	15.3	18.8	21.6	16.7	39.0	21.9	20.7	20.9	17.9	15.8	16.2	18.7	38.5	27.7	36.4	38.8
ChatGpt 3	13.7	12.9	18.0	20.1	17.7	18.7	20.1	20.2	14.3	12.2	21.3	20.7	19.2	20.0	20.2	20.7
ChatGpt 4	17.2	16.4	20.3	17.5	20.0	19.6	21.3	18.0	17.2	16.8	20.7	17.6	20.1	19.0	21.8	19.1
Mpt 7b	24.4	19.8	13.1	22.4	16.0	19.4	nan	nan	25.0	38.7	12.3	16.8	15.9	17.2	nan	nan
Mpt 30B	18.0	13.1	11.4	17.1	15.2	11.1	15.3	15.1	15.2	17.8	10.4	22.8	21.2	22.6	nan	nan
Vic. 13B	13.9	13.5	13.2	13.9	16.2	18.3	15.5	13.2	13.3	15.8	10.4	15.8	16.9	14.0	13.3	9.3
Vic. 33B	11.7	10.3	11.6	12.4	12.8	18.8	13.8	15.1	13.6	15.7	10.0	13.4	16.1	11.4	7.7	5.9
W. V. 13B	12.8	11.6	14.0	14.7	7.3	10.9	17.0	12.1	13.5	10.3	10.9	14.0	8.7	7.9	12.4	9.5
Human	5.1															
Fal. 7B	85.5	45.5	34.0	24.4	22.0	23.7	23.2	26.7	76.6	43.7	26.1	28.6	19.3	23.9	19.1	22.2
Fal. 40B	10.8	16.0	15.2	13.3	33.4	25.6	17.8	23.3	16.6	12.2	11.2	17.0	35.4	32.4	34.9	31.3
ChatGpt 3	4.2	4.6	7.0	6.7	7.1	7.5	8.9	7.4	4.6	5.1	6.5	5.6	10.5	10.4	6.9	6.9
ChatGpt 4	4.2	4.7	4.4	5.6	4.8	4.7	4.8	5.1	5.0	5.5	5.3	5.5	4.6	4.5	4.3	4.7
Mpt 13B	36.9	27.7	10.8	22.3	15.7	18.9	nan	nan	42.7	33.7	12.5	19.3	17.4	19.2	nan	nan
Mpt 30B	29.8	16.0	12.1	16.8	15.5	8.8	12.6	12.5	21.4	17.1	9.9	18.2	20.6	21.1	nan	nan
vic 13B	6.3	10.0	7.7	11.6	15.5	19.0	13.0	13.4	5.4	14.6	8.0	13.9	17.0	15.4	15.3	6.3
Vic. 33B	3.6	3.9	6.6	10.8	15.6	16.7	10.8	12.3	9.1	15.3	11.9	13.2	18.7	15.1	12.1	10.6
W. V. 13B	5.0	10.3	8.4	13.1	9.3	11.8	14.7	9.3	5.5	5.6	9.7	8.7	10.6	12.7	11.1	11.9

Table 4.8: Average lengths (in pinks) and standard deviations (in greens) of the responses from the models for all experiments. On the top are reported the same metrics computed from the crowdsourced eliciting questions as reference. On the rows is reported the model, while in the left is reported the experimental setup, with/without guidelines/colour information for each number of examples (shots) given in the prompt. On the first row of each, the human reference value. Notice how all models are more loquacious than the humans.

Table 4.8 are reported the statistics for average sentence length and standard deviation on the answers of the models, compared to the human respective statistics. Similarly to what happened to the divergence, through these statistics, it is not possible to determine if a model is performing well, but it is possible to exclude models whose statistics do not match the expected ones. From this table, we can observe that many models are more loquacious than human annotators. We can also notice that the eliciting questions from Falcon models, Falcon 7B and Falcon 40B, are particularly long, because they contain non-sense text after a few tokens.

On a deeper layer of inspection, a small investigation on the best examples for each model reveals that at least those examples are indeed good eliciting questions. In Table 4.9 are reported the examples, with ChatGPT 3.5 and 4 reporting elicitations that are almost word by word the same eliciting question as the crowdsourced ones.

Combining everything together we come to understand that the best examples are obtained with guidelines, without colour and in the 5 shot examples.

4.3 Human Evaluation

After the comparison of the models through automatic metrics and manual inspection, it was decided to proceed with a more in depth human evaluation. This evaluation aims to understand if the eliciting questions from LLMs in the ANE task are comparable to crowdsourced ones. In order to do so, the Human Evaluation Protocol presented in [35]

Examples of best eliciting questions from the models			
Model name	**Narrative**	**Model eliciting question**	**Human eliciting questions**
ChatGPT 4	Mi sento leggero. Sono in viaggio per andare a trovare mia figlia.	Oh, che bello! Dove vive tua figlia?	Mi fa piacere, dove abita tua figlia?
ChatGPT 3	Ciao. Oggi giornata abbastanza tranquilla. Non si lavora ma si hanno faccende domestiche e giardinaggio	Che tipo di faccende domestiche devi fare oggi?	Che bello, che faccende devi fare a casa?
Falcon 40B	Mi sento leggero. Sono in viaggio per andare a trovare mia figlia	Come ti senti quando sei in viaggio?	Mi fa piacere, dove abita tua figlia?
Vicuna 33B	Mi è dispiaciuto stamattina non aver avuto pazienza con mio padre. Mi sono resa conto che è davvero difficile, ma ripensandoci dopo, magari invece avrei potuto riuscirci	Sì, capisco. Cosa ti ha fatto perdere pazienza?	Posso capirti. Cosa ti ha fatto perdere la pazienza?
Wiz. Vic.	Mi è dispiaciuto stamattina non aver avuto pazienza con mio padre. Mi sono resa conto che è davvero difficile, ma ripensandoci dopo, magari invece avrei potuto riuscirci	Mi dispiace che tu ti sia sentito male dopo la conversazione. Mi hai spiegato cosa hai appreso da questo episodio?	Posso capirti. Cosa ti ha fatto perdere la pazienza?

Table 4.9: This table reports a few of the best eliciting questions for the models. In the first column the model in question, in the second column the narrative, and in the third column the eliciting question of the model for that specific narrative. The reference crowdsourced eliciting question is reported in the rightmost column. Notice how ChatGPT 3.5 and 4 predict almost word by word the same eliciting questions as crowdworkers. All the examples are obtained with 5 shots as examples, with guidelines and without colour information.

was followed.

It was decided to apply all 4 metrics presented in the Human Evaluation Protocol: Correctness, Appropriateness, Contextualisation and Listening. Aside from Correctness which is related to language, we are particularly interested in models that perform well in Appropriateness, Contextualisation, and Listening. These three metrics can act as proxy for the key requirements of conveying the feeling of emphatetic active listening to the narrator in the ANE task. Compared to the original human evaluation protocol, there is one key difference, which lies in the fact that the evaluation is being applied to narratives instead of dialogues.

Overall, we believe that the 4 metrics presented in the Human Evaluation Protocol can act as good surrogate to understand if a given eliciting question can effectively elicit a continuation of a personal narrative. However, we do recognize that the lack of an ad-hoc metric to measure adherence to the guidelines, with a particular focus on empathy, is a challenge. To address this, in future works, we plan on adding a metric specifically for this issue.

We decided to apply the human evaluation protocol to 4 models, using the eliciting questions gathered from the experimental setting that included 5-shot examples, the use of guidelines, but not the use of colour information as this settings was found to be the overall best performing one. The models selected are:

- Falcon 7B. This model has subpar performance, and it is going to be the lower bound.
- ChatGPT 3.5 turbo. Although ChatGPT 4 should perform better, we ultimately opted for ChatGPT 3.5 turbo as it displayed better behaviour with the increasing number of examples. Furthermore, the token distribution of its generated questions is slightly more similar to the human reference data. With all the analysis done so far, this model is expected to perform extremely well.
- Wizard Vicuna 13B Uncensored HF. This model performed very well compared to others, especially with few shots and considering its smaller size.
- Vicuna 33B V1. Last included model. This model also performs well according to our analysis, although not as well as ChatGPT.

Alongside those 4 models, a 5th data point was added:

- Crowdsourced eliciting questions. This data is going to serve as the upper bound and reference.

For the actual evaluation process, the same UI presented in the Human Evaluation Protocol was used. Three workers were gathered, and they were assigned narratives to be evaluated. In order to prevent annotation fatigue, the narratives were split into batches. Each batch was done daily with an estimated annotation time of 40 minutes. Each of these batches contained a number from 5 to 8 narratives. Each narrative contained the selected 5 elicitation questions: 5 from the models and the human elicitation. In the batches, the first one acted as a training example. For this reason, all narratives for all annotators were exactly the same in this initial batch. This fact allowed the computation of agreement metrics and determined if issues in the annotators' comprehension of the task were present. After this initial training batch, the next batch contained only 2 narratives in common and successive batches only contained 1 narrative in common. In total 4 batches were evaluated.

Overall, this meant that for the 57 narratives that are present in the test set of our corpus, 285 total eliciting questions were annotated. Of those narratives, 9 were evaluated by all three annotators, whereas the remaining 48 narratives were annotated by only one annotator. Considering this redundancy and the fact that there are 4 metrics for each eliciting question, a total of 1500 data points were collected.

Initially, the Fleiss' Kappa [61] metric was calculated for the 5 narratives contained in the first batch as soon as that annotation task was completed. This allowed us to determine that there were no particular issues in the annotators' comprehension of the guidelines for the human evaluation protocol.

Afterwards, the Fleiss' Kappa agreement metric was computed each time for both individual batches and overall metrics. In Table 4.10 are reported the results that were obtained. It is possible to observe that most of the annotators have a high agreement on all metrics except for correctness. Upon further investigation, it was found that the low score is due to the fact that the Fleiss' Kappa metric also accounts for the distributions of the annotations labels. Since most models have correct answers, the very few instances of incorrect values heavily penalise the resulting metric. In order to solve this issue, a decision to compute the percent of agrement was made. This metric accounts for how many answers are shared across annotators. In Table 4.11 are shown the percent of agreement. This metric did show that, in fact, the annotators agree on correctness with a similar ratio to the other metrics measured.

Fleiss' Kappa agreement metric over all annotators						
Batch	**N° of narratives**	**Correctness**	**Appropriateness**	**Contextualisation**	**Listening**	**Global**
Training	5	0.16	0.47	0.61	0.60	0.70
Batch 1	2	0.58	0.47	0.83	0.73	0.80
Batch 2	1	-0.07	0.46	0.55	0.44	0.62
Batch 3	1	0.44	0.76	0.36	1.00	0.81
Overall	9	0.16	0.47	0.61	0.60	0.70

Table 4.10: Fleiss-Kappa agreement metric computed over the different batches for all annotators. On the rightmost column the global agreement computed independently of metric, and on the bottom the overall metrics without accounting for batches. In the second column are also reported the number of narratives that are used to compute the metrics for each batch. Notice how correctness is lower than the other metrics due to the unbalance in distributions of correct and incorrect labels.

% of agreement across all annotators						
Batch	**N° of narratives**	**Correctness**	**Appropriateness**	**Contextualisation**	**Listening**	**Mean**
Training	5	0.60	0.56	0.68	0.64	0.62
Batch 1	2	0.70	0.70	0.90	0.80	0.78
Batch 2	1	0.20	0.60	0.60	0.60	0.50
Batch 3	1	0.60	0.80	0.60	1.00	0.75
Overall	9	0.60	0.56	0.68	0.64	0.62

Table 4.11: Percentage of agreement in the answers computed over the different batches for all annotators. r. On the rightmost column the mean agreement, and on the bottom the overall metrics without accounting for batches. In the second column are also reported the number of narratives that are used to compute the metrics for each batch. Notice that mostly all annotators do agree on their evaluations.

% of votes according to human annotators for the different metrics					
Model	**Cor.**	**App.**	**Con.**	**Lis.**	**Mean**
Human	0.90	0.90	0.86	0.81	0.87
ChatGPT	0.74	0.86	1.00	0.88	0.87
Wiz. Vic. 13B	0.48	0.53	0.66	0.43	0.53
Vicuna 33B V1	0.34	0.43	0.64	0.53	0.49
Falcon 7B	0.50	0.22	0.28	0.16	0.29

Table 4.12: Table reporting the percentage of positive, negative and uncertain values for the 4 models evaluated by 3 human annotators. Each subcolumn represent the evaluation for a different task, Correctness, Appropriateness, Contextualisation, and Listening. A value of 0.90 in Correctness means that for all narrative eliciting questions of that model, 90% of them were evaluated by 3 human annotators as correct. Notice how ChatGPt is very close to human reference.

These two metrics were also computed for each of the three pairs of annotators to highlight with more detail if there were issues in their agreement. It was found that the annotators do mostly agree on their annotations.

With the notion that all annotators mostly agree with each other, highlighting no issues in the understanding of their evaluation task, a scoring system was implemented. The scores were computed for each model, counting the times the eliciting questions proposed by the model were evaluated as Correct, Appropriate, Contextualised, and Listening. The same scoring system was applied for the opposite cases of Incorrect, Inappropriate, Not Contextualised, Not Listening, and Uncertain labels. In the case of eliciting questions annotated by all 3 annotators, the majority vote was taken as valid. These results are shown in Table 4.12. Interestingly, only ChatGPT follows very similar characteristics to humans. All other models are significantly subpar compared to human annotators.

Example of human evaluation results					
Narrative	Ritornata dal lavoro mia figlia mi dice di aver chiamato il medico perché ha dei dolori alla testa, non ha il senso dell'olfatto e del gusto per cui ci siamo un po' allarmati.				
Model name	**Eliciting question**	**Cor.**	**App.**	**Con.**	**Lis.**
Human	Mi dispiace di sentirlo, da quanto tempo tua figlia ha questi sintomi?	✓	✓	✓	**Unsure**
ChatGPT	Mi dispiace sentire che tua figlia sta avendo questi sintomi preoccupanti. Cosa ha detto il medico riguardo ai suoi sintomi?	✓	✓	✓	✓
Wiz. Vic.	Mi dispiace molto, come ti senti ora? Hai avuto altri sintomi? Hai già fatto una visita al medico?	✓	✓	✗	✗
Vic. 33B	Speri che sia solo il stress del lavoro?	✗	✓	✓	✓
Fal. 7B	Sì, è normale dopo una caduta di questo genere. 'NARRATIVA: 'Mia moglie è al telefono con la nostra dottoressa. Ci sono un po' preoccupati perché la testa di mia figlia sanguina. Il medico ha detto però di non preoccuparci. 'DOMANDA: 'Mi dispiace sentirlo, però non si preoccupi tanto. 'NARRATIVA: 'Mia figlia ha la febbre. Le ho fatto il bagno e l'ho messa a letto. Non si è sve	✗	✗	✗	✗

Table 4.13: Examples of eliciting questions of a narrative. On the first row is reported the narrative that is given as context. Following on each row are the models that were evaluated. In the second column are reported the eliciting questions that were evaluated and in the 4 rightmost columns are reported the correct or incorrect values for each of the 4 metrics that were evaluated (Cor. App. Con. Lis.). Consider that even human answers are not 100% accurate according to other annotators, as in this example, the annotator was unsure whether the crowdsourced answer was listening.

It is not surprising to see that human data does not score 1 in all categories, because it was collected through crowdsourcing, and neither the guidelines nor the manual review of the crowdworkers responses considered the criteria present in the Human Evaluation Protocol.

In the Table 4.13 is reported an example of narratives and respective eliciting questions, with their evaluations. It is possible to notice that ChatGPT performs extremely well, whereas Falcon 7B is mostly incoherent. Wizard Vicuna is a surprisingly good model in the examples reported, although it does sometimes have issues with the logical coherence of its output relating to the narrative, which results in low listening scores. Similar reasoning can be applied to Vicuna 33B.

From these results of human evaluation, we can determine that ChatGPT is rated, according to other human annotators, very similarly to crowdworkers according to the Human Evaluation Protocol. Both human and ChatGPT score high in the key categories of Contextualisation and Listening, which act as surrogate to measure the empathy required in eliciting the continuation of a personal narrative, in particular for narrative of sorrowful content.

5 Conclusion

In this work, we presented the *Automatic Narrative Elicitation* (ANE) task, in which a model is required to generate a question that induces the continuation of a personal narrative. Key elements are that the model should convey the feeling of active listening to the topics presented in the narrative, and deliver empathetic responses when required. We also presented a novel corpus designed for this task, which included crowdsourced eliciting questions for continuations of personal narratives. We tested different LLMs on the ANE task using this corpus and we evaluated the results. This evaluation encompassed both automatic metrics and human evaluation.

The human evaluation consisted in a comparison of LLMs generated eliciting questions against crowdsourced ones using the Human Evaluation Protocol [35], which included Correctness, Appropriateness, Contextualisation and Listening. The metrics of Appropriateness, Contextualisation and Listening were used as proxy for the requirement of conveying empathetic responses in the ANE task.

Our findings suggests that LLMs, in particular ChatGPT, manage the grounding context provided to generate adequate correct and contextualised questions based on the narrative topic.

We did find that the many of other smaller models tested did not reach such performances, and were lacking compared to ChatGPT. Many models had a few recurrent issues. Some models recurrently produce either blank or non-valid characters as output. Reprompting might be a solution. A few of the smaller LLMs has a few issues in their abilities to understand and follow topics with common sense reasoning. We suggest that their poor ability to understand common sense may be related to their issues with the Italian language. Curating the training dataset and procedure is a possible solution.

Regarding the performance discrepancy of ChatGPT and other open-source models, is it likely due to two facts. One is the size differences, as ChatGPT has 175B parameters [62] whereas the open-source models tested in this thesis, range from 7B to 40B parameters. The second is that although the ChatGPT pipelines are not transparent, it is very likely that their models include a built-in pipeline which includes a pre and post-processing step in order to clean up the data. However, due to their closed source nature, we cannot confirm this. It should also be noticed that version of ChatGPT tested during this thesis is a paid model, which can be limiting factor on the accessibility and usage of its models.

5.1 Future Works

In future works, we plan on fine-tuning LLMs with the corpus presented in this thesis, focusing on the valence information. The goal is to obtain models that can correctly elicit continuation for personal narratives, with a deeper focus on their Italian language abilities and their ability in prompting questions related to valence, negative or positive functional units of the narratives, as key emphatetic listening behaviour.

Bibliography

[1] *The healing heart for families: storytelling to encourage caring and healthy families*. New Society Publishers, Washington.

[2] Wizard vicuna lm. `https://github.com/melodysdreamj/WizardVicunaLM`.

[3] Learn prompting: Your guide to communicating with ai. `https://learnprompting.org/docs/basics/roles`, 2023.

[4] Hussam Alkaissi and Samy I McFarlane. Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus*, 15(2), 2023.

[5] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

[6] Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):1–2, 2023.

[7] Andrea Bacciu, Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, and Fabrizio Silvestri. Fauno: The italian large language model that will leave you senza parole!, 2023.

[8] Patricia Hill Bailey and Stephen Tilley. Storytelling and the interpretation of meaning in qualitative research. *J. Adv. Nurs.*, 38(6):574–583, June 2002.

[9] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[10] Lee Anne Bell. Telling tales: What stories can teach us about racism. *Race Ethnicity and Education*, 6(1):3–28, 2003.

[11] BootStrap. Bootstrap. `https://getbootstrap.com`.

[12] 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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[13] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[14] Rita Charon. Narrative medicine as witness for the self-telling body. *J. Appl. Commun. Res.*, 37(2):118–131, May 2009.

[15] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. 2018.

[16] Morena Danieli, Tommaso Ciulli, Seyed Mahed Mousavi, Giorgia Silvestri, Simone Barbato, Lorenzo Di Natale, and Giuseppe Riccardi. Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: Randomized controlled trial. *JMIR Ment Health*, 9(9):e38067, Sep 2022.

[17] DeepL. Deepl. `https://www.deepl.com/en/translator`.

[18] Gavin J Fairbairn. Ethics, empathy and storytelling in professional development. *Learn. Health Soc. Care*, 1(1):22–32, March 2002.

[19] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, September 2021.

[20] Google. Material design. `https://m3.material.io`.

[21] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *arXiv e-prints*, pages arXiv–2308, 2023.

[22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv e-prints*, pages arXiv–2009, 2020.

[23] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023.

[24] HuggingFace. Huggingface open-source llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

[25] Sandra Jovchelovitch and Martin W Bauer. Narrative interviewing. In *Qualitative Researching with Text, Image and Sound*, pages 58–74. SAGE Publications Ltd, 6 Bonhill Street, London England EC2A 4PU United Kingdom, 2000.

[26] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[27] Jeong-Hee Kim. *Understanding narrative inquiry*. SAGE Publications, Thousand Oaks, CA, May 2015.

[28] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc., 2022.

[29] Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5(6):e333–e335, 2023.

[30] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[31] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. 2022.

[32] MosaicML. Introducing mpt-7b: A new standard for open-source, commercially usable llms. `https://www.mosaicml.com/blog/mpt-7b`.

[33] MosaicML. Mpt-30b: Raising the bar for open-source foundation models. `https://www.mosaicml.com/blog/mpt-30b`.

[34] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories.

[35] Seyed Mahed Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella, and Giuseppe Riccardi. Evaluation of response generation models: Shouldn't it be shareable and replicable? In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 136–147, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[36] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv e-prints*, pages arXiv–2305, 2023.

[37] Kate P Nurser, Imogen Rushworth, Tom Shakespeare, and Deirdre Williams. Personal storytelling in mental health recovery. *Ment. Health Rev. (Brighton)*, 23(1):25–36, March 2018.

[38] Desmond C Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. Modeling emotion in complex stories: the stanford emotional narratives dataset. *IEEE Transactions on Affective Computing*, 12(3):579–594, 2019.

[39] Desmond C Ong, Zhengxuan Wu, Tan Zhi-Xuan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. Modeling emotion in complex stories: the stanford emotional narratives dataset. *IEEE Trans. Affect. Comput.*, 12(3):579–594, July 2021.

[40] OpenAI. Chatgpt. https://openai.com/chatgpt.

[41] OpenAI. Chatgpt can now see, hear, and speak. `https://openai.com/blog/chatgpt-can-now-see-hear-and-speak`.

[42] OpenAI. Gpt-4 technical report, 2023.

[43] Jonas Oppenlaender. Prompt engineering for text-based generative art. *arXiv e-prints*, pages arXiv–2204, 2022.

[44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[45] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv e-prints*, pages arXiv–2302, 2023.

[46] Prolific. Prolific. `https://www.prolific.co`.

[47] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[48] Laria Reynolds and Kyle McDonell. 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, 2021.

[49] Gabriel Roccabruna, Steve Azzolin, and Giuseppe Riccardi. Multi-source multi-domain sentiment analysis with BERT-based models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 581–589, Marseille, France, June 2022. European Language Resources Association.

[50] Gabriel Roccabruna, Alessandra Cervone, and Giuseppe Riccardi. Multifunctional iso standard dialogue act tagging in italian. 2020.

[51] Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models' strengths and biases.

[52] Amanda L Sammantha and Jessica Vasquez. *Conducting research interviews on sensitive topics*. 2021.

[53] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022.

[54] Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. Annotation of emotion carriers in personal narratives.

[55] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. 2023.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[57] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[59] Lilian Weng. Prompt engineering. https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/.

[60] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv e-prints*, pages arXiv–2302, 2023.

[61] Wikipedia. Fleiss' kappa. `https://en.wikipedia.org/wiki/Fleiss%27_kappa`.

[62] Wikipedia. Gpt-3. `https://en.wikipedia.org/wiki/GPT-3`.

[63] Wikipedia. List of languages by total number of speakers. `https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers`.

[64] Wikipedia. Pearson correlation coefficient. `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`.

[65] Wikipedia. Wasserstein distance. `https://en.wikipedia.org/wiki/Wasserstein_metric`.

[66] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?

[67] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[68] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[69] Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Prompt consistency for zero-shot task generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2613–2626, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[70] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.