Can LLMs Help Recollect and Elaborate On Our Personal Experiences?

Gabriel Roccabruna[1,†], Olha Khomyn[1], Michele Yin[1,*] and Giuseppe Riccardi[1]

[1]*Signals and Interactive Systems Lab,*
Department of Information Engineering and Computer Science, University of Trento

Abstract

In the act of narration, speakers engage with others, communicate findings, and share personal facts and knowledge. This act involves recollecting and reasoning about thoughts and events. Individuals need to plan and organize events and associated emotions in a temporal and logical order. These recollection processes are cognitively demanding and emotion-laden. In this work, we investigate whether Large Language Models (LLMs) may help and support the process of personal narration, i.e. in elaborating on the unfolding events, participants, and emotions. For this, we test LLMs' abilities on a novel task called **A**utomatic **N**arra**T**ive **E**licitation (ANTE). We have crowdsourced a corpus of elicitation responses in the Italian language using a pre-existing dataset of personal narratives. We used this dataset to evaluate a set of closed and open-source LLMs with automatic and human-evaluation metrics. The human evaluation results show that GPT-4 achieves performance similar to humans', while smaller open-source LLMs struggle with this task. We investigate whether fine-tuning smaller open-source LLMs improves performance by experimenting with mixing crowd-sourced and synthetic data.

Keywords

Personal Narrative, Large Language Models, Elicitation, Emotions, Conversational Agent

1. Introduction

The act of narration manifests in written or spoken conversations. It is generally used to communicate facts, knowledge and personal events. This act involves recollecting and reasoning about thoughts and events. Indeed, the narrative has been widely used in journalism [1], education [2], and economics [3]. In psychology, the analysis of personal narratives is a research tool used in many fields such as rehabilitation [4], managing psychosis [5], investigating language dysfunctions [6], and monitoring the variation of the emotional state during psychotherapy [7, 8]. A Personal Narrative (PN) is a series of unfolding events recounting the social interactions, emotions, experiences and others lived by the narrator [9]. In this sense, a PN is a way to observe the interpretation of the world from the narrator's perspective [10, 11].

Currently, the collection of personal narratives is mainly based on textual stimuli or interviews. In the textual stimuli approach, the narrators recount or write down in complete isolation an event [12] recollected by a crafted eliciting prompt based on valence-charged words (e.g. *friendship* or *death*) or questions [13, 14]. However, the act of narration may be a cognitively demanding and

Figure 1: An example of the Automatic NarraTive Elicitation (ANTE) task. A skilled AI agent can help the narrator recall entities and events. Following an opening dialogue act, the model asks a question to support the narrator in continuing, expanding and connecting previous entities, facts, and shared emotions.

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Work done while he was working at the University of Trento.

†Work done prior to joining Amazon.

gabriel.roccabruna@unitn.it (G. Roccabruna); olha.khomyn@unitn.it (O. Khomyn); giuseppe.riccardi@unitn.it (G. Riccardi)

0000-0001-5704-5363 (G. Roccabruna); 0000-0002-0739-8184 (G. Riccardi)

emotionally intense process, leading some individuals to get stuck with the narration or to recount overgeneralized memories, overlooking important details of the story [15, 12]. While human-human conversation has been shown to alleviate these issues [16, 17, 18], the potential role of Large Language Models (LLM) in supporting this process remains underexplored. Indeed, the recent suggested improvements in the safety, biases and toxicity [19, 20] and in natural language fluency [21] make these models suitable candidates for this task.

To help narrators recollect and elaborate on personal events, LLMs must understand the unfolding events, participants, and emotions encompassed in the Personal Narrative (PN). In this work, we investigate whether LLMs have these capabilities by evaluating their performance on a novel task called **A**utomatic **N**arra**T**ive **E**licitation (ANTE). In this, to support the elaboration of personal events, the model is tasked to generate empathetic eliciting responses pointing to a specific aspect of the recount. We crowdsource a corpus of more than 500 eliciting responses in the Italian language starting from a pre-existing dataset of PNs. On this, we evaluate 5 open and closed-source LLMs with in-context learning. The human evaluation has shown that while GPT-4 [22] achieves on-pair performance with the human reference, all the open-source models lag behind. As closed-source LLMs may have privacy issues and not be affordable over the long run, we explore whether fine-tuning small open-source LLMs can reduce the gap. For this, we augment the training set with a partition generated by GPT-4. We then experiment with different combinations of partitions (crowd-sourced vs synthetic data) during fine-tuning. The results show that fine-tuning with synthetic data improves the performance of all models, closing the gap with the human reference.

Our contributions can be summarized as follows:

- Definition of a novel LLM skill for supporting personal narrations;
- Proposed guidelines and procedure for collecting the Automatic Narrative Elicitation (ANTE) corpus;
- Automatic and human evaluation of 5 LLMs following in-context learning and fine-tuning strategies;
- Human evaluation protocol with two task-specific metrics for the ANTE task;

2. Related Works

Question Generation Question Generation (QG) is a natural language processing task in which a model is tasked to generate a question given a context and a target answer [23]. Automatic NarraTive Elicitation (ANTE) is related to QG because the model has to generate a question given a context, but in ANTE the target answer is unknown. Thus, the ANTE task has no predecessors to the best of our knowledge, but previous research in QG is still relevant. GPT-2 [24] has been on the generation of clarifying questions by experimenting with several zero-shot prompts grounding the generation on a list of facets which are possible directions for an ambiguous query [25]. A BART model [26] has been used to generate questions based on a storybook summary for improving intellectual development in children [27]. In the healthcare domain, a combination of T5 and BERT models has been used in the task of asking patients with depression questions for triage [28].

Data Augmentation with LLMs Recently, there has been increased attention on using the LLMs for data augmentation. [29] have leveraged several LLMs to augment three multilingual datasets. Similarly, [30] have developed an augmentation method based on GPT-3 [31] and in-context learning to generate a dataset of synthetic dialogues. Related to this, the ability of LLMs to generate Socratic questions, i.e. questions for helping students solve a problem without revealing the answer, has been investigated [32]. For this, the authors augmented a dataset with GPT-4 [22] and fine-tuned Llama2 [33] with reinforcement learning.

3. Automatic Narrative Elicitation

We envision a hybrid methodology for eliciting Personal Narratives (PN), which joins the benefits of textual stimuli and interview approaches. The elicitation, depicted in Figure 1, starts with an eliciting prompt such as a crafted textual stimulus. Then, once the narrator finishes the first part of the recount an agent asks a follow-up response that helps continue the narration by elaborating on some aspect of the story. These exchanges go on till a certain criterion is met, depending on the application (e.g. based on the narrative length), or the narrator explicitly wants to stop.

Formally, a prompt P elicits the main event of the PN. This is followed by a sequence $d = [(N_1, R_1), ..., (N_t, R_t)]$, where N_t is a narrative turn at time t and R_t is the corresponding eliciting response. R_t consists of feedback and an eliciting question. The feedback must show active listening and be aligned with the expressed narrator's emotions. Furthermore, the eliciting response must focus on relevant events mentioned in N_t $(from\ 1\ to\ n)$ without significantly altering the flow of the narration.

The Automatic NarraTive Elicitation (ANTE) task is defined as:

Definition 3.1. *Given the sequence* $[(N_1, R_1), ..., N_t]$*, the model generates a* R_t *such that* R_t *elicits the narrator to continue with the story by yielding a* N_{t+1}*.*

This task implicitly requires an emotional and semantic understanding of the narrative. Furthermore, it implicitly requires the ability to select the events that might be valuable to support the continuation of the narration.

4. Data Collection

The dataset for the ANTE task has been created starting from an existing dataset of PNs in the Italian language collected during a psychological study, CoAdapt [34]. This corpus is composed of PNs about daily experiences collected from 45 subjects suffering from distress and stress conditions. The vocabulary size of the corpus is 3355 words, showing a wide semantic diversity of the narratives. Furthermore, the PNs are annotated with valence and emotion carriers at the functional unit level [35]. The functional unit is a concept borrowed from the dialogue act theory, which identifies the minimum span of text with a communicative function [36, 37].

To collect the ANTE corpus, we have asked the annotators to write an eliciting response based on a personal narrative taken from the CoAdapt corpus. Due to data and resource constraints, we collected only one response for each narrative. Thus, the evaluation is based only on the generation of the first response. With a model trained on this task, a possible solution to enable the collection of multi-turn dialogues is using an adaptation of the Wizard-of-Oz framework [38], in which the system could be a machine or a human supported by a machine. This would reduce the complexity and costs of the data collection.

To guide the annotators in writing eliciting responses aligned with our definition, we have written a list of hints to follow, which is:

- **Focus on the narrative**: the response has to be focused on emotionally charged (*preferred*) or other events mentioned in the narrative;
- **Give feedback**: the response should contain a feedback signal or other signs of active listening;
- **Show empathy**: the response should be empathetic, i.e. showing understanding of the emotions expressed;
- **Be short**: the response should be brief and to the point;

Similarly, we have included the description of undesirable properties, such as asking for personal opinions, and hypothetical events, giving suggestions or shifting the focus of the conversation away from the narrated event. Furthermore, to help the annotator focus the question on emotionally charged events, we have included the valence values by highlighting with red and green colours positive and negative functional units, respectively. The web interface and the guidelines are available on GitHub[1], to foster the reproducibility of the data collection.

Table 1
Statistics of the CoAdapt corpus and the eliciting responses in the ANTE dataset composed of the Crowdsourced, Merged and Synthetic datasets.

	Crowds.	Merged	Synth.
# Narratives	478	478	478
# Elicit. Resp.	561	897	478
AVG Tok. Resp.	12.1	15.4	18.6
Vocab. Size Resp.	1061	1878	1350

The annotators have been hired through the Prolific platform[2]. Only Italian native speakers who passed a qualifying test have been considered eligible for this task to ensure data quality. The CoAdapt dataset has been split into batches of seven narratives each to keep control of the cognitive load by keeping the duration of each annotation session below 20 minutes. Each batch has been assigned to five crowd workers. As an additional quality check, we have used an overlap of 20%, which has been inspected manually. We have kept this overlap in the training set to have more training data and removed it from the test set by random sampling one of the eliciting responses. We set the compensation for the workers to £12 per hour.

Additionally, to train open-source LLMs we have augmented this corpus using GPT-4. For each narrative in the dataset, an eliciting response is generated using the API[3] provided by OpenAI. The prompt given to the model is presented in Section 5.1.

Overall, we have used three datasets to evaluate the models on the ANTE task: (i) Crowdsourced, containing only human-annotated responses (ii) Synthetic, containing only GPT-4 eliciting responses (iii) Merged, containing both human-annotated and GPT-4 generated eliciting responses. Table 1 reports the statistics of the datasets, in which the number of eliciting responses for the crowdsourced dataset is higher than the synthetic dataset due to the overlap. We used the official data split of the CoAdapt corpus.

5. Methods

We have experimented with 5 closed-sourced and open-source LLMs, namely GPT-4, Llama3 8B [39], Vicuna 13B [40], LLaMAntino 13B [41], and IT5 [42]. The

[1] https://github.com/sislab-unitn/ANTE
[2] https://www.prolific.com/
[3] We used `gpt-4-turbo`

selection of the models has only considered LLMs supporting the Italian language i.e. the language of the ANTE dataset. IT5 is pre-trained on the Italian dataset, while LLaMAntino 13B based on Llama2 [33] is fine-tuned on the Italian language using LoRa [43]. Instead, Llama3 8B and Vicuna 13B are pre-trained on a multi-lingual dataset.

5.1. In-Context Learning

In-context learning, or few-shot learning, is a technique in which the model can learn from a few examples provided in the context [31]. In our case, five pairs (5-shot) of narratives and corresponding eliciting responses are given to the model. In particular, we have used the same examples written in the guidelines for collecting the dataset.

The input to the model is formalized as:

$$I \oplus \{N_1^1, R_1^1 \oplus ... \oplus N_1^5, R_1^5\} \oplus N$$

where I are the instructions for the model, $\oplus$ is the concatenation with the new line (\n), N_1^i, R_1^i are i-shot example of the narrative and the corresponding eliciting response at the first turn of the dialogue, N is the input narrative that the model should generate the response to. The beginning of the narrative and the response are indicated with two marker tokens, namely "*Narrative:*" and "*Response:*" [4]. We have also experimented with adding the annotation guidelines before the instructions for the model, but observed only an increase in inference time and not in performance.

5.2. Fine-tuning

In training, the input sequences consist of a narrative and the corresponding eliciting response, concatenated with the new line (\n). Additionally, we add two marker tokens to the input prompt to indicate the beginning of the narrative and the response, respectively.

Formally, the input sequence is:

$$Narrative : N \oplus Response : R$$

where N is the narrative, $\oplus$ is the concatenation with the new line and R is the corresponding eliciting response. In fine-tuning the open-source LLMs, the input of the autoregressive models is as described above, while for the sequence-to-sequence IT5 model, the input to the encoder and decoder is narrative and eliciting response, respectively. All the hyperparameters used to fine-tune and test the models are reported in Appendix A.

[4] An example of a real prompt is reported in Appendix A in Table 5.

6. Evaluation

6.1. Metrics

We have evaluated the models on the ANTE task both with automatic and human evaluation metrics. We have used the automatic metric to have a proxy for performance estimates during the development of the models, i.e. before the resource-demanding human evaluation. As an automatic evaluation metric, we have used the BLEU 1 score [44]. Regarding the human evaluation, we have adopted a human evaluation protocol developed for evaluating dialogue models in a reproducible and comparable way [45]. From this, we have used the *Appropriateness*, *Contextualization* and *Correctness* metrics[5]. Each metric is translated into a question to which the annotators can answer *Yes*, *No*, or *I don't know*. Furthermore, the annotators can provide explanations for a negative answer for some metrics. For *contextualization*, the annotators can justify their negative answer with *wrong* or *no references* to the grounding context representing hallucination and genericness, respectively.

While the proposed metrics are enough for evaluating generic dialogue models, we need specific criteria for better evaluating the models on our task. Specifically, we introduced *Effectiveness* and *Compliance*. *Effectiveness* evaluates whether the response is effective in helping the narrator continue with the narration naturally. The two possible explanations for being an ineffective response are that the question is either generic (*generic question*) or complex (*complex question*), which means the narrators will have difficulties in answering that question. Different from the *genericness* in *contextualization*, a generic response can still be effective when the context is not enough for asking a more specific question. *Compliance* evaluates whether the response is compliant with the annotation guidelines, i.e. it has the properties listed in Section 4.

Additionally, in the HE, we have added ground truth eliciting responses along with those generated as a point of reference and an additional control step [45]. Moreover, as for the data collection, we have split the evaluations into batches of five narratives. Each batch has been annotated by five crowd workers hired via Prolific and paid £9 per hour. Furthermore, we used an overlap of 20% to compute the agreement, whose overall score is 0.34 measured with Fleiss'κ [46], showing a fair agreement.

[5] *Appropriateness* whether the response makes sense w.r.t the dialogue history; *Contextualization* whether the response contains references to the dialogue context; *Correct* whether the response is grammatically and syntactically correct.

Table 2
The table reports the BLEU 1 scores for each model tested on *Gold* and *Silver*, i.e. the crowdsourced and synthetic test sets, respectively. We can observe that Vicuna 13B and IT5 fine-tuned on a crowdsourced dataset achieve better results than GPT-4. Moreover, Llama3 8B fine-tuned on the synthetic dataset outperforms all the other models on the *Silver* test set.

	GPT-4	**Llama3 8B**		**Vicuna 13B**		**LLaMAntino 13B**		**IT5**	
	Gold	*Gold*	*Silver*	*Gold*	*Silver*	*Gold*	*Silver*	*Gold*	*Silver*
ICL	0.15	0.06	0.07	0.09	0.12	0.09	0.07	0.08	0.10
Crowdsourced	-	0.15	0.11	**0.16**	0.13	0.14	0.13	**0.16**	0.19
Merged	-	0.14	0.18	0.09	0.13	0.13	0.17	0.13	0.12
Synthetic	-	0.12	**0.22**	0.12	0.16	0.10	0.17	0.12	0.16

Table 3
Human evaluation results achieved with in-context learning (ICL) and fine-tuning (FT) on the *Crowdsourced* (Crowds.), *Merged* and *Synthetic* corpora. The results on the left of || are given to facilitate the comparison. In ICL, GPT-4 outperforms all the other models, matching human performance in most of the metrics. Open-source models achieve the highest performance when synthetic data is added to fine-tuning (*Merged* and *Synthetic* rows). Yet all the models have a significant gap in the compliance metric, but Llama3 fine-tuned on the *Synthetic* corpus.

	Metrics	**Human Ref.**	**GPT-4**	**Llama3 8B**	**Vicuna 13B**	**LLaMAn. 13B**	**IT5**
ICL	Appropriateness	90.2	**90.2**	29.4	54.9	60.8	5.9
	Contextualization	96.1	**98.0**	27.5	64.7	66.7	9.8
	Correctness	94.1	**94.1**	41.2	62.7	92.2	29.4
	Compliance	90.2	**80.4**	31.4	64.7	76.5	19.6
	Effectiveness	96.1	**92.2**	31.4	70.6	70.6	17.6
FT. Crowds.	Appropriateness	90.2	90.2	59.3	45.1	58.8	11.8
	Contextualization	96.1	98.0	68.5	62.7	58.8	35.3
	Correctness	94.1	94.1	94.4	66.7	80.4	52.9
	Compliance	90.2	80.4	81.5	62.7	64.7	62.7
	Effectiveness	96.1	92.2	72.2	60.8	66.7	23.5
FT. Merged	Appropriateness	90.2	90.2	70.6	60.8	66.7	27.5
	Contextualization	96.1	98.0	74.5	74.5	76.5	45.1
	Correctness	94.1	94.1	68.6	52.9	66.7	62.7
	Compliance	90.2	80.4	78.4	82.4	82.4	82.4
	Effectiveness	96.1	92.2	82.4	76.5	86.3	49.0
FT. Synthetic	Appropriateness	90.2	90.2	**84.3**	52.9	78.4	13.0
	Contextualization	96.1	98.0	**86.3**	64.7	78.4	22.2
	Correctness	94.1	94.1	**94.1**	66.7	92.2	50.0
	Compliance	90.2	80.4	**88.2**	74.5	76.5	72.2
	Effectiveness	96.1	92.2	**92.2**	68.6	90.2	35.2

6.2. Automatic Evaluation

Table 2 reports the BLEU 1 score for each model attained with in-context learning and fine-tuning on crowdsourced, merged and synthetic datasets. As ground truth, we use both gold and silver eliciting responses coming from the crowdsourced and synthetic test sets, respectively.

From the results of the in-context learning experiments, we observe that GPT-4 outperforms all the other models by effectively leveraging the provided examples with few shots. Fine-tuned on the crowdsourced dataset, Vicuna 13B and IT5 outperform GPT-4 with ICL, achieving the highest results on the gold test set overall. Furthermore, while fine-tuning the models on the merged and synthetic datasets always degrades the performance measured on the gold test set, it generally increases the scores on the silver test set. Finally, Llama3 8B fine-tuned on the synthetic dataset achieves the best BLEU score on the silver test set.

According to these results, Llama3 8B and IT5 should have similar performance on the ANTE task. Notwithstanding, recent studies have shown that automatic metrics are poorly correlated with human judgement [47, 48, 45]. For this reason, we have used human evaluation to have a more realistic representation of the LLMs' performance.

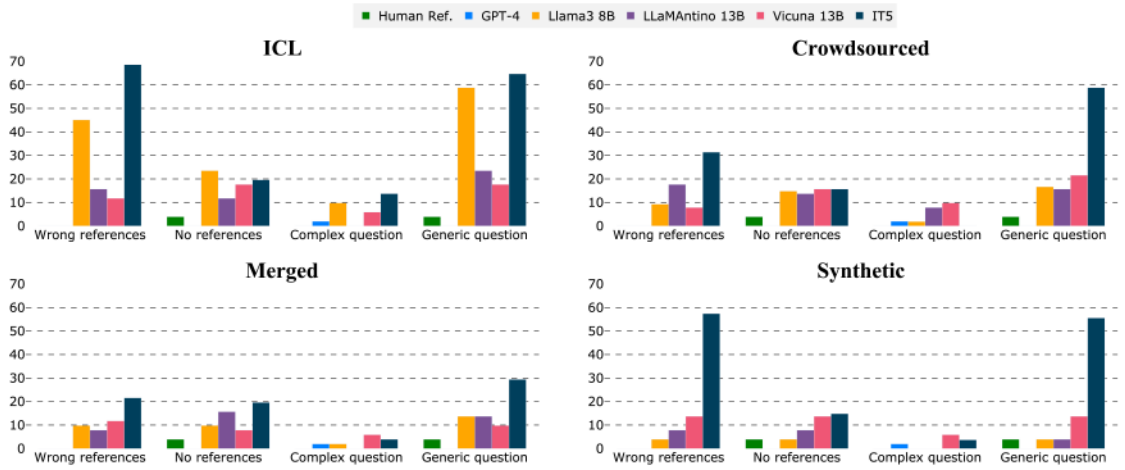

Figure 2: The figure depicts the percentages of the errors classified by annotators for the metrics Contextualization (*Wrong* and *No references)* and Effectiveness (*Complex* and *Generic questions*). We can observe that the errors are mainly due to hallucinations and genericness, which are minimized by adding synthetic data to fine-tuning.

6.3. Human Evaluation

The results of the human evaluation are presented in Table 3. Similarly to the automatic evaluation, the table shows the results achieved with ICL and fine-tuning on crowdsourced, merged and synthetic datasets. The values represent the percentage of eliciting responses that received a positive evaluation for the corresponding metric. Considering the limited size of the test set (57 examples) and the unavoidable subjectivity and ambiguity in the evaluation process, the results are compared with a coarse margin that we empirically set to ± 5. Along with manual inspection, this is also supported by the percentage of "*I don't know*" options, catching the ambiguous cases, which ranges from 3.5% for human reference to 9.1% for Vicuna 13B on average.

The results in the ICL setting show that the ANTE task is challenging also for crowd workers (*human reference*) who in some cases could not refrain from giving suggestions or asking for personal information (e.g. *What's the name of your kid?*). Moreover, GPT-4 achieves on-pair performance with human annotators on all metrics but *compliance* since the model gave suggestions similar to the human reference. Given the overall positive scores, we have used GPT-4 to generate the synthetic data. Regarding the other models, the gap with human reference is overall large. Only LLaMAntino 13B and Vicuna 13B achieve decent performance on the two task-specific metrics *compliance* and *effectiveness*. Moreover, the scores on *correctness* suggest that only LLaMAntino 13B and GPT-4 can properly handle the Italian language in this task without fine-tuning.

Fine-tuning especially boosts the performance of IT5 and Llama3 8B, while more contained improvements are observed for LLaMAntino 13B and Vicuna 13B. Moreover, LLaMAntino 13B and Llama3 8B achieve their best results when fine-tuned on the synthetic dataset, whilst IT5 and Vicuna 13B perform the best when fine-tuned on the merged dataset. In particular, Llama3 8B fine-tuned on the synthetic dataset attains an improvement of 35% on average w.r.t. ICL results, outperforming all the other open-source LLMs and matching the performance on the task-specific metrics of human annotators and GPT-4. Although a lower performance gain, 10% on average, LLaMAntino 13B is the second-best model on the ANTE task, matching GPT-4 performance on *effectiveness* and *correctness*. Regarding the *correctness* metric, we can observe that IT5 always achieves the lowest score, but on the merged dataset, despite being pre-trained on a corpus in the Italian language.

All in all, fine-tuning with synthetic data (either merged or synthetic datasets) improves the performance of almost all the models. Indeed, the scores of the task-specific metrics achieved by fine-tuning the models on the crowdsourced dataset are lower on average than those achieved with merged and synthetic datasets. A possible explanation for these improvements is that the merged dataset is larger; therefore, a small model such as IT5 (220M parameters) benefits from this.

6.4. Error Analysis

Since the human evaluation has shown that GPT-4 matches the Human Reference's (HR) performance, we have run some analysis to characterize the similarities and differences better. We have started by manually com-

Table 4
Entrainment statistics computed between the eliciting responses in test sets (crowdsourced and synthetic) and the eliciting responses generated by two best-performing fine-tuned models. The score is defined between 0 (*perfect match*) and -1 (*mismatch*).

Test sets	**Fine-tuned on Crowdsourced (FTC)**		**Fine-tuned on Synthetic (FTS)**	
	Llama3 8B	LLaMAntino 13B.	Llama3 8B	LLaMAntino 13B
Crowdsourced (CT)	-0.52	-0.55	-0.58	-0.54
Synthetic (ST)	-0.66	-0.60	-0.46	-0.35

paring the eliciting responses of GPT-4 and HR. In this, we observed that GPT-4 tends to use paraphrased parts of the narrative in the feedback and question parts of the eliciting response. Indeed, the Jaccard similarity [49] between the narrative and the eliciting response[6] on average is 13% for GPT-4 and 7% for HR. After that, we investigate whether there is a challenging set of examples on which both models make errors by considering an eliciting response wrong when it received negative feedback on at least one metric. The intersection of the errors is only the 7% of the narrative, while the cases in which HR is correct and GPT-4 is wrong are 20% and vice versa are 13%. By analysing all these errors manually, we observed that in some cases GPT-4 deducted the context wrongly such as "*I was having a coffee with a colleague and we were talking about Christmas when...*" and the model asked[7] "*Have you already decided what to gift for Christmas?*". Overall, one of the main issues is due to suggestions or requests for personal information negatively affecting the performance on *appropriateness* and *compliance*.

The distributions of the explanations that annotators gave to justify their negative evaluations for the metrics *contextualization* (*wrong* and *no* references) and *effectiveness* (*complex* or *generic* questions) are depicted in Figure 2. HR and GPT-4 errors are reported as references in all plots. We can observe that HR is penalized on *contextualization* and *effectiveness* due to genericness in the responses. On the GPT-4 side, the negative score on *effectiveness* is mainly due to complex questions. Furthermore, the percentage of errors classified as *wrong references* is zero for both HR and GPT-4, meaning that GPT-4 does not hallucinate in this task. The opposite is observed in the ICL experiments where Llama3 8B has been penalized on *contextualization* mainly due to wrong references, i.e., the model hallucinated some part of the eliciting response. Moreover, for the same model, the *effectiveness* score is negatively affected by many generic questions. As for human evaluation, the distributions of the errors show that fine-tuning the models improves the performance, especially with synthetic data. In this, we can observe that the cases of hallucination and genericness on the synthetic dataset are minimized compared to fine-tuning on the crowdsourced dataset. The improvement is even more evident comparing the errors of IT5 fine-tuned on crowdsourced and merged datasets, where the number of generic questions is halved, and the hallucination cases decrease by 11%. All in all, we can observe that the major source of errors for *contextualization* and *effectiveness* is due to either hallucination or genericness, regardless of the dataset used during fine-tuning.

We have investigated whether the performance gap between fine-tuning on crowdsourced and synthetic datasets is due to a difference in the learning complexity. In other words, learning from synthetic data may be easier than learning from human-generated data. Our rationale is that the distribution learned by LLMs, during pre-training, is more similar to the distribution of synthetic data than that of human-generated data. This is because LLMs are based on similar architectures, and the relative pre-training datasets may overlap. For this, we have used the entrainment statistic because of the different vocabularies, making measuring the distribution distance challenging. Entrainment is the phenomenon in which, during a conversation, a speaker reuses the terms of the other interlocutor [50]. This phenomenon may also be seen during the training process, where a model learns to use the same language as the training set. We have measured the entrainment using the formula proposed by Hirschberg et al. [51], which is:

$$ENTR(c) = -\frac{\sum_{w \in c} |count_{S_1}(w) - count_{S_2}(w)|}{\sum_{w \in c} |count_{S_1}(w) + count_{S_2}(w)|} \quad (1)$$

where c is a target word class and $count_{S_i}$ is the frequency of the word w used by the model S_1 and the test set responses S_2. The resulting score ranges between 0 (*perfect match*) and -1 (*mismatch*). We used the 100 most frequent words computed on the joint responses generated by S_1 and S_2.

Specifically, as S_1, we have used the responses generated by either Llama3 8B or LLaMAntino 13B[8] fine-tuned on crowdsourced (FTC) and synthetic (FTS) datasets. As S_2, we have used the responses either in the crowdsourced (CT) or the synthetic (ST) test sets. From Table

[6] From both, we removed the stopwords and lemmatized the rest.
[7] In this case, the model wrongly inferred that Christmas is yet to come, which is impossible to say by looking at the context only. The model should have focused on other parts of the narrative.
[8] The two best-performing models.

4, we can observe that the entrainment scores computed between FTC and CT are lower than those computed between FTS and ST. Thus, the fine-tuned models are more aligned with the language of the synthetic dataset than the natural language found in the crowdsourced dataset, suggesting that learning from the synthetic data is easier.

7. Personal Narratives in VR

To test the models in a real-case scenario, we have developed a Virtual Reality (VR) system for the collection of personal narratives. The collection follows the same procedure as depicted in Figure 1, which starts with an eliciting prompt and is followed by a conversation between a narrator and an embodied conversational agent. The system consists of an automatic speech recognition [52] model, a conversational agent based on our best-performing LLM (Llama3 8B), which generates eliciting responses, and a text-to-speech model. To connect these components, we have utilized an adaptation of the architecture proposed by Yin et al. [53], which also employs a strategy of input segmentation to minimize response latency. After some internal tests, we have observed that the dialogue is effective and the system's response latency is not a major issue. However, the turn-taking strategy is rule-based and, therefore, studying a more effective approach would make the conversation smoother[9].

8. Conclusions

In this work, we evaluated 5 LLMs on the Automatic NarraTive Elicitation (ANTE) task to investigate whether the models can help us elaborate and recollect personal events. To do this, we collected and created three corpora, namely crowdsourced, merged, and synthetic. Then, we evaluated closed and open-source models with in-context learning and fine-tuning on the ANTE task. The results show that closed-source LLMs can perform similarly to human annotators and that fine-tuned open-source LLMs on synthetic data can achieve similar performance. This suggests that LLMs may be used to support individuals in recollecting and elaborating on personal events.

A future work is to study the effectiveness of LLMs in collecting personal narratives compared to standard techniques such as textual stimuli or interviews in a random controlled trial setting. Another is to study how to instruct the model to steer the conversation toward specific events relevant to the researchers or professionals collecting the narratives.

[9]A demo of this system can be found at https://www.youtube.com/watch?v=ozpuoEKsTjs

Acknowledgments

We acknowledge the support of the MUR PNRR project FAIR - Future AI Research (PE00000013) and the MUR PNRR project iNEST- Interconnected Nord-Est Innovation Ecosystem (ECS00000043) funded by the European Union under NextGenerationEU. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or The European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] T. B. Connery, A sourcebook of american literary journalism: representative writers in an emerging genre (1992).

[2] L. Hobbs, R. Davis, Narrative pedagogies in science, mathematics and technology, Res. Sci. Educ. 43 (2013) 1289–1305.

[3] R. J. Shiller, Narrative economics: How stories go viral and drive major economic events, Princeton University Press, 2020.

[4] K. D'Cruz, J. Douglas, T. Serry, Personal narrative approaches in rehabilitation following traumatic brain injury: A synthesis of qualitative research, Neuropsychological Rehabilitation 29 (2019) 985–1004.

[5] C. N. Wiesepape, J. T. Lysaker, S. E. Queller, P. H. Lysaker, Personal narratives and the pursuit of purpose and possibility in psychosis: directions for developing recovery-oriented treatments, Expert Review of Neurotherapeutics 23 (2023) 525–534.

[6] N. Botting, Narrative as a tool for the assessment of linguistic and pragmatic impairments, Child language teaching and therapy 18 (2002) 1–21.

[7] M. Danieli, T. Ciulli, S. M. Mousavi, G. Silvestri, S. Barbato, L. Di Natale, G. Riccardi, Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: randomized controlled trial, JMIR mental health 9 (2022) e38067.

[8] G. Roccabruna, S. M. Mousavi, G. Riccardi, Understanding emotion valence is a joint deep learning task, in: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, 2023, pp. 85–95.

[9] A. Tammewar, A. Cervone, E.-M. Messner, G. Riccardi, Annotation of emotion carriers in personal narratives, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France,

2020, pp. 1517–1525. URL: https://aclanthology.org/2020.lrec-1.189.

[10] T. R. Sarbin, The narrative as a root metaphor for psychology, Narrative psychology: The storied nature of human conduct (1986) 1–27.

[11] U. Neisser, R. Fivush, The remembering self: Construction and accuracy in the self-narrative, 6, Cambridge University Press, 1994.

[12] C. Mills, S. D'Mello, On the validity of the autobiographical emotional memory task for emotion induction, PloS one 9 (2014) e95837.

[13] J. M. Williams, K. Broadbent, Autobiographical memory in suicide attempters., Journal of abnormal psychology 95 (1986) 144.

[14] D. C. Rubin, Remembering our past: Studies in autobiographical memory, Cambridge University Press, 1999.

[15] R. J. McNally, N. B. Lasko, M. L. Macklin, R. K. Pitman, Autobiographical memory disturbance in combat-related posttraumatic stress disorder, Behaviour research and therapy 33 (1995) 619–630.

[16] G. Borrini, P. Dall'Ora, S. Della Sala, L. Marinelli, H. Spinnler, Autobiographical memory. sensitivity to age and education of a standardized enquiry, Psychological Medicine 19 (1989) 215–224.

[17] M. D. Kopelman, B. Wilson, A. D. Baddeley, The autobiographical memory interview: a new assessment of autobiographical and personal semantic memory in amnesic patients, Journal of clinical and experimental neuropsychology 11 (1989) 724–744.

[18] B. Levine, E. Svoboda, J. F. Hay, G. Winocur, M. Moscovitch, Aging and autobiographical memory: dissociating episodic from semantic retrieval., Psychology and aging 17 (2002) 677.

[19] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, et al., Llama guard: Llm-based input-output safeguard for human-ai conversations, arXiv preprint arXiv:2312.06674 (2023).

[20] T. Rebedea, R. Dinu, M. N. Sreedhar, C. Parisien, J. Cohen, Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2023, pp. 431–445.

[21] J. Ou, J. Lu, C. Liu, Y. Tang, F. Zhang, D. Zhang, K. Gai, DialogBench: Evaluating LLMs as human-like dialogue systems, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 6137–6170. URL: https://aclanthology.org/2024.naacl-long.341. doi:10.18653/v1/2024.naacl-long.341.

[22] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[23] J. Qiu, D. Xiong, Generating highly relevant questions, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5983–5987. URL: https://aclanthology.org/D19-1614. doi:10.18653/v1/D19-1614.

[24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[25] Z. Wang, Y. Tu, C. Rosset, N. Craswell, M. Wu, Q. Ai, Zero-shot clarifying question generation for conversational search, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 3288–3298.

[26] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[27] Z. Zhao, Y. Hou, D. Wang, M. Yu, C. Liu, X. Ma, Educational question generation of children storybooks via question type distribution learning and event-centric summarization, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5073–5085. URL: https://aclanthology.org/2022.acl-long.348. doi:10.18653/v1/2022.acl-long.348.

[28] S. Gupta, A. Agarwal, M. Gaur, K. Roy, V. Narayanan, P. Kumaraguru, A. Sheth, Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts, in: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, 2022, p. 137.

[29] C. Whitehouse, M. Choudhury, A. F. Aji, LLM-powered data augmentation for enhanced cross-lingual performance, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computa-

tional Linguistics, Singapore, 2023, pp. 671–686. URL: https://aclanthology.org/2023.emnlp-main.44. doi:10.18653/v1/2023.emnlp-main.44.

[30] Z. Li, W. Chen, S. Li, H. Wang, J. Qian, X. Yan, Controllable dialogue simulation with in-context learning, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4330–4347. URL: https://aclanthology.org/2022.findings-emnlp.318. doi:10.18653/v1/2022.findings-emnlp.318.

[31] T. Brown, B. Mann, R. et al., Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[32] N. Ashok Kumar, A. Lan, Improving socratic question generation using data augmentation and preference optimization, in: E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, Z. Yuan (Eds.), Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 108–118. URL: https://aclanthology.org/2024.bea-1.10.

[33] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[34] S. M. Mousavi, A. Cervone, M. Danieli, G. Riccardi, Would you like to tell me more? generating a corpus of psychotherapy dialogues, in: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, Association for Computational Linguistics, Online, 2021, pp. 1–9. URL: https://aclanthology.org/2021.nlpmc-1.1. doi:10.18653/v1/2021.nlpmc-1.1.

[35] S. M. Mousavi, G. Roccabruna, A. Tammewar, S. Azzolin, G. Riccardi, Can emotion carriers explain automatic sentiment prediction? a study on personal narratives, in: Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 62–70. URL: https://aclanthology.org/2022.wassa-1.6. doi:10.18653/v1/2022.wassa-1.6.

[36] H. Bunt, V. Petukhova, D. Traum, J. Alexandersson, Dialogue act annotation with the iso 24617-2 standard, in: Multimodal interaction with W3C standards, Springer, 2017, pp. 109–135.

[37] G. Roccabruna, A. Cervone, G. Riccardi, Multifunctional iso standard dialogue act tagging in italian, in: CLiC-it, 2020.

[38] J., F. E. Kelley, T. J. Watson, An iterative design methodology for user-friendly natural language office information applications, ACM Trans. Inf. Syst. 2 (1984) 26–41. URL: https://api.semanticscholar.org/CorpusID:207660078.

[39] A. Grattafiori, A. Dubey, A. J. et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[40] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

[41] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, 2023. URL: https://arxiv.org/abs/2312.09993. arXiv:2312.09993.

[42] G. Sarti, M. Nissim, IT5: Text-to-text pretraining for Italian language understanding and generation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 9422–9433. URL: https://aclanthology.org/2024.lrec-main.823.

[43] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[44] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[45] S. M. Mousavi, G. Roccabruna, M. Lorandi, S. Caldarella, G. Riccardi, Evaluation of response generation models: Shouldn't it be shareable and replicable?, in: A. Bosselut, K. Chandu, K. Dhole, V. Gangal, S. Gehrmann, Y. Jernite, J. Novikova, L. Perez-Beltrachini (Eds.), Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 136–147. URL: https://aclanthology.org/2022.gem-1.12. doi:10.18653/

v1/2022.gem-1.12.
[46] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological bulletin 76 (1971) 378.
[47] A. Belz, S. Mille, D. M. Howcroft, Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing, in: Proceedings of the 13th International Conference on Natural Language Generation, Association for Computational Linguistics, Dublin, Ireland, 2020, pp. 183–194. URL: https://aclanthology.org/2020.inlg-1.24.
[48] A. B. Sai, A. K. Mohankumar, M. M. Khapra, A survey of evaluation metrics used for nlg systems, ACM Computing Surveys (CSUR) 55 (2022) 1–39.
[49] P. Jaccard, Nouvelles recherches sur la distribution florale, Bull. Soc. Vaud. Sci. Nat. 44 (1908) 223–270.
[50] S. E. Brennan, et al., Lexical entrainment in spontaneous dialog, Proceedings of ISSD 96 (1996) 41–44.
[51] J. B. Hirschberg, A. Nenkova, A. Gravano, High frequency word entrainment in spoken dialogue (2008).
[52] J. Grosman, Fine-tuned XLSR-53 large model for speech recognition in Italian, https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-italian, 2021.
[53] M. Yin, G. Roccabruna, A. Azad, G. Riccardi, Let's give a voice to conversational agents in virtual reality, in: Proceedings of Interspeech 2023, Dublin, Ireland, 2023, pp. 5247–5248.
[54] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. URL: https://arxiv.org/abs/1412.6980. arXiv:1412.6980.
[55] N. Shazeer, M. Stern, Adafactor: Adaptive learning rates with sublinear memory cost, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 4596–4604. URL: https://proceedings.mlr.press/v80/shazeer18a.html.

A. Appendix

A.1. Hyperparameters

We used a batch size of 8 for the fine-tuning. The models were fine-tuned for 10 epochs with early stopping based on the perplexity computed on the development set. We have trained the autoregressive models, Vicuna 13B, LLaMAntino 13B, Llama3 8B, in an auto-regressive manner with Adam [54] optimizer. The models were fine-tuned using Low-Rank Adaptation (LoRA) [43], i.e. a method for fine-tuning large-scale LLMs, which reduces the number of trainable parameters. We set the learning rate to $1e{-}5$, rank and alpha parameters to 128. We have used the top-k sampling strategy to generate the new tokens with k set to 10. The IT5 model was fully fine-tuned with Adafactor [55] optimizer. We have used a beam search with four beams as a decoding strategy. To run our experiments, we used a machine with two Nvidia 3090 with 24GB and an Nvidia A100 with 80GB. Overall, the training time for each experiment was less than 30 minutes, and the inference time was less than 15 minutes.

Table 5
This is the prompt we have used in the in-context learning experiments and to generate the synthetic dataset with GPT-4. The prompt that we used is in Italian. In the second row, we provide a translated version.

Sei una AI che deve generare una risposta empatica con 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 su cui dovrai generare una risposta con 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."
RISPOSTA: "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."
RISPOSTA: "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." RISPOSTA: "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."
RISPOSTA: "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."
RISPOSTA: "Mi dispiace tanto, da quanto erano insieme?"
NARRATIVA: 'input narrative'
RISPOSTA:

You are an AI that has to generate an empathic response with a question about a story to get more information about events that happened in the story. Below are some examples followed by a narrative, on which you will have to generate a response with a question to get more information.
NARRATIVE: "Today was a beautiful day. My wife told me that she is expecting a baby! I am super happy! I wonder if I will be a good father. My father was not very present when I was a child."
RESPONSE: "I am happy to hear that. Do you already know if it is a boy or a girl?" NARRATIVE: "Today I argued with Chiara, she was angry with me because in her opinion I don't know how to do things."
RESPONSE: "Oh, I am sorry that you argued. What does she think you don't know how to do?"
NARRATIVE: "Today is a beautiful day. I went ice skating and then I went to the cinema." RESPONSE: "It is nice to hear that it was a good day for you. Where did you go skating?"
NARRATIVE: "I was always thinking about my son who had to go out in the afternoon, this is the reason that triggered my anxiety."
RESPONSE: "I understand, where was your son supposed to go?"
NARRATIVE: "My daughter broke up with her boyfriend, and now I feel guilty and sad, I'm so sorry, and I feel unable to support her in this. In short, somewhat gray days. I don't know if the disturbed sleep and some episodes of insomnia are caused by this confusion."
RESPONSE: "I'm so sorry, how long were they together?"
NARRATIVE: 'input narrative'
RESPONSE: