Can we use LLMs to generate data-driven patient narratives? A multi-method exploration in the oncology domain

Nele Albers[*1], **Linwei He**[*1], **Anouk Teunissen**[2], **Kelly de Ligt**[2], **Nadine Bol**[1], **Emiel Krahmer**[1]
[1]Tilburg University, Tilburg, Netherlands, [2]Netherlands Cancer Institute, Amsterdam, Netherlands
{n.albers, l.he_1, nadine.bol, e.j.krahmer}@tilburguniversity.edu, {a.teunissen, k.d.ligt}@nki.nl

Abstract

Translating patient data about health-related quality of life (e.g., fatigue) to narrative feedback has the potential to enhance the interpretability and emotional resonance of key outcomes in oncology care. However, producing personalized narratives at scale is challenging. Based on advances in large language models (LLMs), this study explores whether LLMs have the potential to generate narratives that are of sufficiently high accuracy and quality to support patients' health-related quality of life. Therefore, we generated 5500 narratives from 110 oncology patient profiles using 10 LLMs varying in size, domain specificity, and reasoning capabilities as well as 5 prompt paraphrases. Using a human validated LLM-as-judge approach, we find that LLMs generate narratives of varying accuracy, with the best model achieving over 80% in accuracy. Human evaluations show that most narratives are seen as accurately following the prompt and as coherent, but vary more for empathy and engagement. Accuracy is overall not associated with empathy or coherence and even weakly negatively associated with engagement, highlighting the importance of multi-criteria evaluation. Further research should explore how narratives can be generated in a robust way in a clinical setting and how clinicians can be meaningfully involved.

1 Introduction

The generation of narratives has long been a topic of research in the field of natural language generation (NLG) (Gatt and Krahmer, 2018), but until recently the generated narratives tended to be relatively simple and straightforward. Large language models (LLMs) bring new possibilities in generating fluent and engaging texts. Thus, previous work has examined if and how LLMs can produce narratives that are coherent, logical, and relevant, including both short- and long-form stories (Gurung and Lapata, 2025; Ismayilzada et al., 2025), and how such narratives can be evaluated using both automatic metrics and human judgments (Chhun et al., 2022; Zhao et al., 2023). Yet, much of this research is situated in the context of creative or literary story generation, and evaluation has typically focused on aspects such as novelty, coherence, and engagement. While these studies provide valuable insights into the generation of high-quality narratives, they place less emphasis on whether the generated content faithfully represents underlying structured information, such as attributes of the characters or events. This raises the question of whether LLMs can generate narratives that are simultaneously coherent and engaging as well as an accurate reflection of such input data. This challenge is particularly relevant in domains where narratives must be grounded in structured data, such as communicating about patient-reported outcome measures (PROMs) in healthcare.

PROMs are validated questionnaires that capture multiple dimensions of patients' experiences, including daily functioning and physical symptoms (Krogsgaard et al., 2021). The resulting scores provide structured insights into patients' health status and are increasingly used to guide clinical decision-making and monitor quality of life during and after treatment (Ishaque et al., 2019). Importantly, PROMs are only effective when results are fed back to and sufficiently discussed with patients (Gibbons et al., 2021). Yet, PROM results are commonly presented as numerical scores or graphical summaries (Reading Turchioe et al., 2020; van der Willik et al., 2021), which are difficult for patients to interpret and may not adequately convey the implications for self-management of their health. One promising approach to addressing this challenge is the use of narrative feedback, which translates structured PROM results into natural-language descriptions that contextualize patients' responses and acknowledge their experiences (Kreuter et al., 2007). Re-

[*]Authors contributed equally.

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 257–294
October 17-21, 2026 ©2026 Association for Computational Linguistics

search in health communication has shown that narrative formats can enhance the interpretability, emotional resonance, and perceived recognition of patient experiences compared to purely numerical feedback (Boomstra et al., 2024; Shaffer et al., 2018). Such narratives may thus help patients better understand their PROM results and support reflection on their health status. However, generic narratives may fail to capture the experience of an individual patient, highlighting the need for personalized narratives. Producing narratives for each individual PROM profile, however, presents substantial practical challenges. Manually creating individualized narrative is labor intensive, making it difficult to implement at scale in routine healthcare settings. As PROM collection becomes increasingly implemented, scalable approaches to generating personalized narrative feedback are needed.

LLMs offer a potential solution to this challenge by enabling the automatic generation of personalized narratives. Although LLMs have demonstrated capabilities in generating creative stories, generating patient-facing narratives based on individual PROM data presents a different challenge. Each narrative must accurately represent specific patterns in the PROM scores, such as elevated symptom burden or preserved functioning, while maintaining coherence and readability. Moreover, these narratives describe aspects of patients' health and well-being, making the communicative context more sensitive than typical narrative generation tasks. Previous research has emphasized the importance of accuracy and faithfulness when using language models in the healthcare domain where generated text reflects underlying data or evidence (Nazi and Peng, 2024; Xie et al., 2023). In the context of patient communication, inaccuracies or misleading interpretations of PROM scores could lead to misinterpretation about patients' health status. At the same time, narratives must remain coherent, supportive, and engaging for emotional resonance.

We thus explore whether LLMs have the potential to generate narratives based on PROM data that are both accurate and of sufficient narrative quality to be meaningful to readers. To this end, we generate a large corpus of narratives from ten LLMs using a diverse set of patient PROM profiles (Figure 1). We then evaluate the generated narratives along two dimensions. First, we assess the accuracy with which the narratives reflect the underlying PROM scores using a human-validated LLM-as-judges approach. Second, we conduct a large-scale human evaluation in which participants rate the perceived quality of the narratives. Through these explorations, we aim to provide first insights into the capabilities and limitations of current LLMs for generating personalized narrative from PROM profiles. Importantly, we do not currently intend or evaluate these narratives as a deployable clinical tool. Instead, we first need to explore how LLMs fare on this task. Future work should then investigate how the narratives can be further improved to eventually support the meaningful interpretation of PROM results in a clinical setting. We discuss requirements for this at the end of this paper.

2 Related work

LLMs for creating narratives. While narrative generation has a rich tradition in the pre-LLM era (Gatt and Krahmer, 2018), approaches tended to be complex (e.g., using a refinement search planning algorithm to obtain sound and believable sequences of story character actions (Riedl and Young, 2010)), hand-crafted for stories with very specific characteristics (e.g., centered around betrayal (Bringsjord and Ferruci, 1999)), and largely capable of producing only short stories (Gatt and Krahmer, 2018). In contrast, LLMs potentially offer a scalable way to generate also longer stories, either independently or by assisting authors (Teleki et al., 2025). However, while some LLMs can independently generate fictional narratives that meet various quality criteria, they are still often outperformed by experienced human writers, for example because they use formulaic or cliched plots or language (Chakrabarty et al., 2024). Besides generating short fiction writing, LLMs have also been used to create plausible stories of people as health interventions. For instance, Bhattacharjee et al. (2025) used LLMs to create tailored narratives acknowledging and addressing challenging thoughts and situations faced by young adults. And Progga et al. (2024) explored creating empathetic and relatable stories tailored to mothers facing postpartum depression. These applications suggest that LLM-generated narratives have the potential to support emotional engagement in health contexts, even though evaluations are often based on automated metrics such as BLEU or ROUGE that tend to be insufficient to capture higher-level narrative quality (Chhun et al., 2024; Teleki et al., 2025). In many of these applications, LLMs do not directly interact with the people they are writing about, but LLMs have also been used

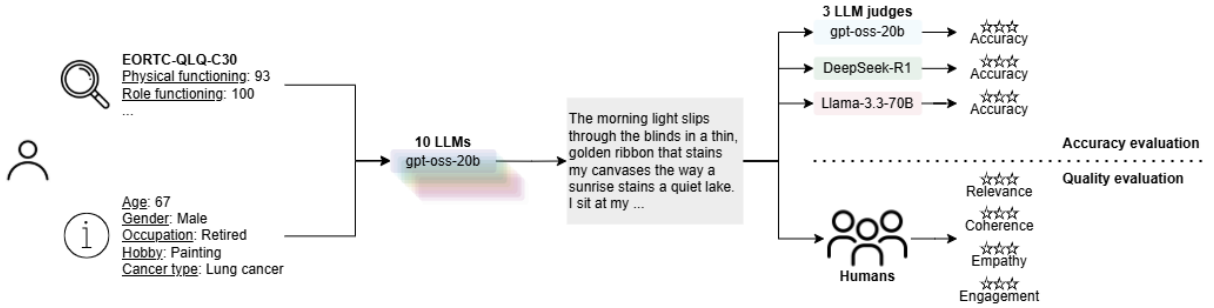


Figure 1: Overview of our approach. A patient's PROM scores and personal characteristics are given to an LLM which is asked to create a narrative describing the patient's experiences. This narrative is subsequently evaluated regarding accuracy by 3 LLM judges and regarding quality by humans.

in interactive settings to create narratives describing people's lived experiences, for example to collect rich datasets that highlight the perspectives of a target group with minimal participant burden (Skeggs et al., 2025). What makes these health applications different from work on fiction writing is that besides the writing quality, the accuracy of the presented information is also crucial. Yet, while previous work has warned about the danger of fake or ambiguous resources in such narratives (e.g., Progga et al. (2024)), limited work has evaluated accuracy and narrative quality simultaneously. This is the focus of this paper.

Narrative communication in cancer care. While prior work has examined how LLMs can generate narratives with varying degrees of quality, their value ultimately depends on how well they address communication needs in specific domains. In oncology, communication is essential in supporting patients' understanding, decision-making, and coping (Albrecht et al., 2008), yet is challenged by complex and abstract medical information and emotional distress (Burg et al., 2015). These challenges are particularly evident for PROMs, where raw numbers are often difficult to interpret and integrate into communication (Boomstra et al., 2024; Thestrup Hansen et al., 2020). Research shows that communication around PROMs is often insufficiently tailored to patients' diverse skills, preferences, and capacities (de Ligt et al., 2025; Westendorp et al., 2023) and therefore requires balancing informational accuracy with emotional responsiveness (Lai et al., 2022). One approach to convey such complex and emotionally laden information are narratives, which have been shown to improve comprehension, recall, and perceived support (Kreuter et al., 2007; Murphy et al., 2013). In health contexts, they can foster identification, emotional engagement, and support mental well-being (Green and Brock, 2000; Bhattacharjee et al., 2025; Rennick-Egglestone et al., 2019). Narratives also provide vicarious experiences, allowing individuals to observe how others cope with similar situations, thereby strengthening self-efficacy (Bandura, 2004; McQueen et al., 2011). In oncology and in communicating about PROMs, however, the use of narratives remains underexplored, with a few studies showing that patients find personalized narratives understandable, meaningful, and supportive (Bekker et al., 2013; Boomstra et al., 2024). Existing approaches typically rely on manual writing or predefined templates, limiting scalability and adaptability to diverse patient characteristics and evolving data.

Building on insights from these two domains, we explore the potential of LLMs in scalably generating narratives from individual PROMs data, and evaluate both their fidelity to the underlying health data and their perceived narrative quality.

3 Methods

Below we describe how we generated the narratives and how we evaluated them regarding accuracy and quality. Our evaluation was preregistered in the Open Science Framework (He et al., 2026).

3.1 Narrative generation

To obtain an LLM-generated narrative for a set of PROM scores, we provided language models with scores from the EORTC QLQ-C30, a cancer patient-reported outcome questionnaire consisting of subscales measuring functioning and symptoms (Aaronson et al., 1993). Each subscale was scored 0–100, with respective thresholds for clinical importance indicating score levels at which symptoms or functional limitations are considered clinically

meaningful (Giesinger et al., 2020). The models were then prompted to generate a narrative describing the patient's experience with the given scores.

Prompt. The prompt consisted of a task description, an explanation of the subscales of the EORTC QLQ-C30 and their thresholds together with the patient's PROM scores, and stylistic guidelines based on Murphy et al. (2013). The prompt structure was designed following prompt engineering practices that recommend explicitly specifying task instructions, contextual information, and output constraints to guide LLM behavior (White et al., 2023). Moreover, we empirically observed that adding personal patient characteristics led to richer narratives. Therefore, the prompt also contained a patient's gender, age, and cancer type following Boomstra et al. (2024) as well as their occupation and hobby. These were sampled from a plausible age range and cancer types, occupations, and hobbies collected from kanker.nl, a Dutch online forum for cancer patients and survivors. And lastly, since small prompt changes can lead to drastic changes in model outputs (Mizrahi et al., 2024; Sclar et al., 2024), we performed all experiments with five prompt paraphrases. We created these by asking ChatGPT based on GPT-5 to rephrase the initial prompt while keeping the same meaning and subsequently correcting any inaccuracies. Figure A1 shows the full prompt.

PROM scores. We generated narratives for 100 fictitious PROM profiles, which we created by randomly sampling a value either below or above the subscales' thresholds for clinical importance. Moreover, we received 10 anonymized real profiles from the Netherlands Cancer Institute (Table A1). The corresponding patients gave consent for the secondary use of their data and anonymized sharing of data with third parties; data is only shared after formal evaluation by the Institutional Review Board (IRB24-335).

Models. We generated 5500 narratives with ten diverse open-weight models differing in size, domain specificity, and reasoning capability: `Llama-3.2-3B-Instruct`, `Llama-3.1-8B-Instruct`, `Llama-3.3-70B-Instruct` (Meta AI, 2024), `Bio-Medical-Llama-3-8B` (ContactDoctor, 2024), `Mistral-7B-Instruct-v0.3` (Mistral AI, 2023), `Command-r-v01` (Cohere, 2024), `DeepSeek-R1-Distill-Qwen-32B` (Guo et al., 2025), `s1.1-32B` (Muennighoff et al., 2025), and `gpt-oss-20b` with both medium and low reasoning levels (OpenAI, 2025). Using Llama models of various sizes thereby allowed us to see if and how model size impacts performance. We focused on open-weight models because of cost and reproducibility (e.g., Dietrich and Hollstein (2025)) as well as privacy concerns. The latter matters as patients in a future clinical setting may not necessarily consent to their PROM profiles being shared externally, even if they are anonymized and hence not identifiable as the 10 real PROM profiles we worked with. For reproducibility, all models were called with the same parameters (temperature = 0.6, top_p = 0.9) and with a random seed.

3.2 Accuracy evaluation with LLMs as judges

To evaluate how accurately the generated narratives represented the PROM scores, we first took an LLM-as-judge approach, which has shown promise as evaluation approach for both general NLP tasks (Bavaresco et al., 2025) and healthcare settings (Li et al., 2026). Specifically, following the work by Boomstra et al. (2024) on personalizing narratives based on whether scores are below or above the threshold, we prompted `gpt-oss-20b`, `DeepSeek-R1-Distill-Qwen-32B`, and `Llama-3.3-70B-Instruct` to indicate for each generated narrative whether the subscale scores were described as below or above the respective threshold for clinical importance (Giesinger et al., 2020) or absent. Following the guidelines by Bavaresco et al. (2025) to compare LLM annotations to human annotations, we chose these three models as judges as they showed the highest agreement with a human judge on a separate set of 108 narratives generated based on random PROM scores among nine tested models. For each judge, we computed the accuracy by comparing the ratings to the PROM scores the narrative was generated for. Subscales rated as absent were considered as incorrectly presented in the narratives. To determine the agreement between the judges, we computed pairwise Spearman correlations between the average accuracy scores per narrative as well as Fleiss κ between the ratings per subscale. We obtained pairwise Spearman correlations of between 0.86 and 0.87 and an average Fleiss κ of 0.72 (Table B1), which suggests substantial agreement (Landis and Koch, 1977). The evaluation prompt is shown in Figure A2.

3.3 Quality evaluation with human participants

Since our goal was to develop narratives that are used to communicate information to humans, we

supplemented our LLM-as-judge evaluation with a large-scale human one. Specifically, to evaluate the perceived quality of the generated narratives, we conducted a human evaluation study with lay participants for which we obtained ethical approval by the Research Ethics and Data Management Committee of the Tilburg School of Humanities and Digital Sciences, belonging to Tilburg University, with code REDC2025.77. Lay participants were recruited because the evaluation focused on general narrative qualities, which have been shown to be reliably assessed by non-expert readers in human evaluations of generated text (Belz and Reiter, 2006; van der Lee et al., 2019). The evaluation was conducted on a stratified subset of 500 narratives, sampled to cover all 10 LLMs and 5 prompt paraphrases, and to include narratives generated from both 5 real PROM profiles and 5 fictitious ones.

Participants were recruited through the Prolific crowdsourcing platform and were required to be fluent in English and at least 18 years old. In total, 300 participants completed the evaluation. These participants were 47.7% female, with an average age of 33, residing in 32 different countries. Each participant evaluated five narratives, and each narrative was rated by three different participants. Narrative assignment was arranged such that no participant saw the same narrative more than once.

Participants were given one narrative at a time and asked to evaluate it along four quality dimensions: relevance, coherence, empathy, and engagement. The mean completion time for evaluating all five narratives was 27.29 minutes (SD = 11.51). The quality criteria were adapted from prior work on human evaluation of automatically generated stories (Chhun et al., 2022) and similar ones have been used in other human evaluations of LLM-generated stories in healthcare (Progga and Rubya, 2026). **Relevance** captures how well the narrative follows the prompt in describing the PROM profile. **Coherence** assesses whether the narrative makes sense and is logically structured. **Empathy** evaluates whether the reader recognizes and understands the character's emotional states, and **engagement** measures whether the reader finds the narrative engaging and interesting. All ratings were collected on ordinal 5-point scales, where 1 indicated the lowest quality and 5 the highest quality for the respective dimension. The formulations of the questions posed to participants can be found in our study preregistration (He et al., 2026).

4 Results

4.1 What do the generated narratives look like?

Figure C1 depicts wordclouds of the narratives. They show that besides mentioning the EORTC QLQ-C30 subscales, the models describe the day of a cancer patient together with emotions and experiences, mentioning both positive and negative aspects. First sentences thereby tend to describe waking up in the morning while last sentences mention resilience and not giving up. In terms of length, there is quite some variation both between and within models (Table C1). For example, the narratives by `s1.1-32B` contain on average 904.65 (SD = 252.09) words, while those generated by `Bio-Medical-Llama-3-8B` contain only 432.82 (SD = 144.12). N-gram diversity across all narratives thereby is rather low (Table C2), for example when compared to human-written news articles (Shaib et al., 2025). Examples of narratives generated by different models are given in Table C3 and all narratives can be found in the online repository accompanying this paper.

4.2 How accurate are the generated narratives?

Figure 2 shows the accuracy scores for each of the ten LLMs, with `s1.1-32B` (M = 0.82, SD across paraphrases = 0.02) being on average most accurate and `Command-r-v01` and `Llama-3.2-3B-Instruct` (M = 0.43, SD = 0.03 and 0.04) being on average least accurate (Table D1). To see how statistically meaningful the model differences are, we performed a Bayesian statistical analysis using the `brms` package by Bürkner (2017)[1]. Comparing two Bayesian multilevel models with and without a random intercept for the ten LLMs shows that the model with the random intercept for the ten LLMs has a lower Widely Applicable Information Criterion (WAIC) (-5131.4) than the model without the random intercept (-1368.1), suggesting that there is a meaningful difference in accuracy between the ten LLMs. Bayesian pairwise comparisons further show that it is virtually certain that the accuracy of `s1.1-32B` is higher than those of all other models (Table D2).

Accuracy thereby varies widely between EORTC QLQ-C30 subscales (Table D3). For ex-

[1]For an introduction to Bayesian statistics see McElreath (2020) and for an overview of advantages of the Bayesian approach see Wagenmakers et al. (2018).

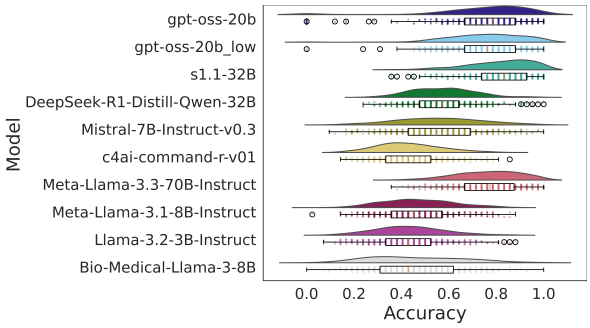


Figure 2: Raincloud plots of the accuracy for each of the ten LLMs. Each raincloud plot consists of a half violin plot, the scattered accuracy scores, and a boxplot.

ample, while the average accuracy across the ten LLMs is 0.76 (SD across paraphrases = 0.02) for social functioning, it is only 0.35 (SD = 0.08) for diarrhea. We suspected that the thresholds could play a role in this, with some subscales having more extreme thresholds than others. To assess the impact that proximity of PROM scores to the threshold may have on the accuracy, we fit a multilevel Bayesian model with a general intercept, random intercepts for the five paraphrases, ten LLMs, and 14 subscales, and a main effect for whether the PROM scores were moderate (i.e., within 5% of the threshold) or extreme (i.e., within the lowest or highest 5% of the scale). We obtained a mean of 0.06 (95%-HDI = [0.05, 0.07], Prob(b $>$ 0) $>$ 0.99995) for the fixed effect, which can be evaluated as virtually certain that extreme scores are represented more accurately than moderate scores based on the interpretation guidelines by Chechile (2020). The corresponding effect size has a probability of 0.99 to be small and one of 0 to be large based on the guidelines by Funder and Ozer (2019).

We also found relatively high variation in accuracy between the five prompt paraphrases for each of the 110 profiles (Figure 3), suggesting that accuracy is quite susceptible to small prompt changes. Moreover, to see if there is a difference in accuracy between the real and fictitious profiles, we fit a multi-level Bayesian model with a general intercept, random intercepts for the five paraphrases and ten models, and a main effect of whether the profiles were fictitious or real. We obtained a value of 0.07 (95%-HDI = [0.06, 0.09], Prob(b $>$ 0) $>$ 0.99995) for the fixed effect, which can be evaluated as virtually certain that real profiles are represented more accurately than fictitious ones (Chechile, 2020). The corresponding effect size has a probability of 0.90 of being large (Funder and

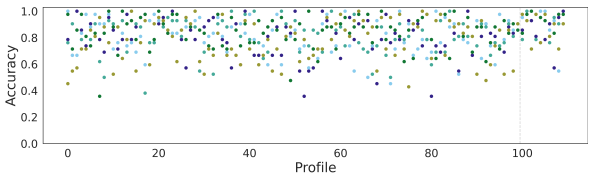


Figure 3: Accuracy for the 5 prompt paraphrases for each of the 110 PROM profiles for `s1.1-32B`. The colors differ per prompt paraphrase. The vertical lines separate the 100 fictitious from the 10 real PROM profiles. Data for more models is shown in Figure D1.

Ozer, 2019). One plausible reason for this could be that our real profiles tend to have extreme scores (Table A1).

4.3 What is the perceived quality of the narratives?

Figure 4 shows the percentage distribution of ratings (1–5) aggregated across all observations for each quality dimension. Across models, the distributions are mainly on the positive side of the scale, indicating generally favorable evaluation of the narratives. Because the ratings were collected on an ordinal scale, we report below the percentage distribution of rounded narrative-level ratings rather than means. Ratings from the 3 raters were averaged and rounded to the nearest integer.

Overall, narratives received especially positive evaluations for coherence and relevance. For coherence, several models received a majority of ratings at the highest scale level. In particular, `Llama-3.3-70B-Instruct` received 52% of level-5 ratings, while `DeepSeek-R1-DistillQwen-32B` and `Llama-3.2-3B-Instruct` received 42% and 44%, respectively. Across models, the large majority of coherence ratings fell within levels 4 and 5. A similar pattern arose for relevance, where most ratings fell at level 4. `s1.1-32B` obtained the largest share of highest relevance ratings, with 32% of narratives rated at level 5 and 60% at level 4.

Ratings for empathy were somewhat more distributed but still largely positive, though with fewer narratives receiving the maximum score. Instead, ratings were more concentrated at level 4, indicating that narratives were generally perceived as empathetic even when not judged as exceptionally so. For example, `Llama-3.3-70B-Instruct` and `Llama-3.2-3B-Instruct` received 66% of ratings at level 4, while the highest share of level-5 ratings was observed for `Llama-3.1-8B-Instruct` and `s1.1-32B` (both 26%).

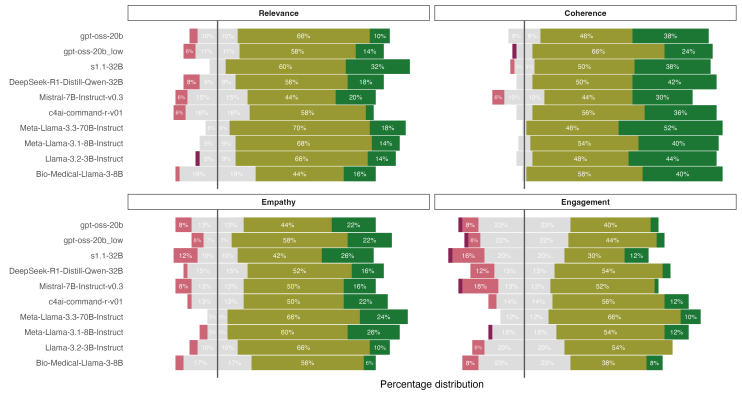


Figure 4: Percentage distribution of quality ratings.

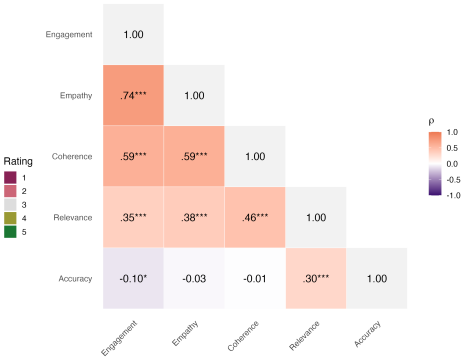


Figure 5: Spearman correlations between accuracy and quality ratings.

In contrast, engagement received lower ratings and showed greater variability across models. While `Llama-3.3-70B-Instruct` again performed relatively well with 66% of ratings at level 4 and 10% at level 5, several models exhibited larger shares of moderate ratings. For instance, `gpt-oss-20b` received 46% of ratings at level 3, indicating that narratives were often perceived as moderately engaging.

Statistical analysis of quality differences between LLMs. To statistically determine if there is a quality difference between the LLMs, we fit two Bayesian multi-level models for each of the four quality measures. The first model contained a general intercept and random intercepts for the paraphrases and PROM profiles, and the second model additionally contained random intercepts for the ten LLMs. Comparing the two models shows that the model with the random intercept for the ten LLMs always obtained a lower WAIC than the model without that random intercept (Table E1), suggesting that there is a difference between the ten LLMs for each quality measures. Bayesian pairwise model comparisons further provide credible evidence that the model with the highest average score is better than most other models for all quality measures except coherence (Table E2 – Table E5).

Statistical analysis of quality differences between real and fictitious profiles. We also used Bayesian multi-level models to assess whether real PROM profiles received higher quality scores than fictitious ones. For each of the four quality measures, we fit a model that contains a general intercept, random intercepts for the paraphrases, ten LLMs, and profiles, and a fixed effect describing whether the profiles are real. We find some evidence that real profiles are rated with lower relevance, empathy, and engagement than fictitious profiles, but little evidence for a difference between profile types for coherence (Table E6).

4.4 Relation between accuracy and quality

Lastly, we computed Spearman rank-order correlation between accuracy and perceived quality, as well as correlations among the four quality dimensions (Figure 5). Each narrative's three ratings for a quality dimension were averaged to obtain a single score to relate to the accuracy score. The four quality dimensions were moderately to strongly correlated with each other, ranging from $\rho = .35$ (relevance-engagement) to $\rho = .74$ (engagement-empathy), with engagement also strongly associated with coherence ($\rho = .59$), suggesting that narratives perceived as coherent and emotionally resonant were also more likely to be engaging. In contrast, accuracy showed little relationship with perceived quality. The strongest positive association was observed with relevance ($\rho = .30$), suggesting that narratives judged as more relevant were also somewhat more likely to accurately reflect the PROM information. Accuracy was not significantly associated with empathy or coherence, and showed a weak negative association with engagement ($\rho = -.10$). Overall, the results suggest that accuracy and perceived quality capture partly related but largely distinct aspects of the narratives. The narratives scoring overall highest and lowest across both quality and accuracy can be found in Table E7 and Table E8, respectively.

5 Discussion and conclusion

In this work we generated 5500 narratives describing cancer patient's experiences using 10 diverse LLMs and 5 prompt paraphrases. Results show that there are substantial differences in estimated accuracy between models, with some (e.g.,

s1.1-32B, Llama-3.3-70B-Instruct) appearing to generate narratives of relatively high accuracy. While Llama-3.3-70B-Instruct also scored relatively high for quality, s1.1-32B obtained among the lowest scores for empathy and engagement, suggesting that accuracy and quality can but do not always go hand in hand. While accuracy was not associated with empathy or coherence, it showed a moderate positive association with relevance and a weak negative association with engagement. The negative association with engagement may also indicate that emphasizing complete and precise information can come at the expense of narrative engagement. Participant feedback suggested that this may be driven by strong instruction-following behavior, where some models prioritize covering all PROM elements in a structured manner, resulting in narratives perceived as mechanical, formulaic, or resembling symptom checklists rather than immersive stories. Potentially, more engaging narratives could thus be created by giving examples of immersive stories or even by fine-tuning LLMs for story generation, which has hardly been investigated (Liu et al., 2026). Other approaches such as explicit instructions on engaging emotional trajectories (Bharti and Sharma, 2026) are also worthwhile to explore. In line with work by Badawi et al. (2026), which shows that high safety and interpretation scores for mental health support messages do not necessarily imply high scores in more affective dimensions such as empathy, our results highlight the importance of multi-criteria LLM evaluation.

Notably, for almost all models and evaluation criteria, there were some narratives that received low or even the lowest possible scores. Together with the relatively high variance in accuracy and quality between different prompt paraphrases (Figure D1, Figure E1 – Figure E4), this illustrates that our narratives are not yet safe to use in a stand-alone application. This is because while narratives could be automatically judged regarding accuracy and re-generated if needed, it is not clear how well the quality dimensions could be automatically scored. Chhun et al. (2024), for example, show that while LLM quality ratings tend to correlate with human ratings similarly or even better than automatic measures, the correlations are still rather low. If one were to further study using our narratives in a clinical context, promising might be a setting where a clinician is present. When used in conversations with clinicians, narratives could potentially help address concerns of cancer patients and their caregivers regarding lack of tailored information and of feeling seen and heard (Westendorp et al., 2023).

However, even for a setting in which clinicians are also present, further improvements to the narratives are necessary. One aspect is that our clinical oncology collaborators noted that for PROM scores worse than the thresholds of clinical importance, more fine-grained accuracy assessments may be needed. For example, to distinguish fatigued patients who cannot leave their bed anymore from those who are severely impacted but still mobile. Moreover, given that our evaluations were only done by LLM judges and lay participants, future work should evaluate the narratives regarding clinical suitability both in general and specifically for use in patient-clinician communication. Besides a more in-depth analysis of accuracy, additional criteria such as informativeness, clinical appropriateness, and actionability may become important. For instance, how accurately do people interpret their PROM results based on our narratives? And how does this compare to interpretation based on numeric PROM scores? Such questions require evaluation by healthcare professionals and ultimately patients themselves. In addition, our prompting strategy included stylistic instructions derived from narrative theory, which may not fully reflect how patients naturally describe their experiences. This could also be culturally sensitive and thus perceived differently by our participants, which were residing in 32 different countries. Future work could thus incorporate patient-generated data, such as interviews or testimonials, to enable more authentic and realistic narratives while also accounting for potential cultural differences. In light of our mixed findings regarding real versus fictitious PROM profiles – where we showed that real profiles led to slightly more accurate but less relevant, empathetic and engaging narratives – real profiles should ideally be used for this.

Besides concerns about robustness and clinical suitability, there are also ones regarding cultural and demographic biases. LLMs tend to be less representative of people with non-western cultural backgrounds (e.g., Naous et al. (2024)), which is especially concerning in a health context given that people from different cultures speak about health and illness differently (Pendse et al., 2019). Moreover, LLMs can have gender biases (e.g., Bajaj et al. (2024)) and can be less representative of people with a lower socioeconomic position or racial minorities (e.g., Kim and Lee (2024)). The latter is

especially problematic when those are also the people who would benefit most from narrative PROM feedback as they are not well served by traditional PROM feedback (Calvert et al., 2022). To generate narratives well suited for a patient at hand, just providing personal characteristics in the prompt like we did is therefore likely not sufficient. Possible improvements could be obtained by fine-tuning smaller models on data from a specific target group, for example as obtained from interviews, patient portals, or public survey data (e.g., Li et al. (2024)). Such personalization must be approached carefully, however, as an uncanny effect could arise if narratives resemble patients too closely (Bhattacharjee et al., 2025). Personalization could also be improved by not presenting all PROMs but only those that patients are concerned about. For example, patients could decide if and how a narrative continues (e.g., Bhattacharjee et al. (2022)). Currently, narratives accurately describing all PROM scores can be rather long (e.g., see the narratives in Table E7).

Overall, we show, based on 5500 narratives generated by ten diverse LLMs for five prompt paraphrases, that some LLMs can generate narratives with relatively high scores for both accuracy and the quality dimensions of relevance, coherence, empathy, and engagement. This supports investigating how these narratives can be further improved to be suitable for clinical use. To facilitate such future work, we publish our code and data.

Ethical considerations

Even though the current paper is dedicated to exploring the potential of LLMs for generating data-driven patient narratives, and not to developing a new clinical application, we do want to discuss a number of ethical considerations that may arise when thinking about the latter.

Making clear that narratives are not describing real experiences. Besides the concerns regarding cultural and demographic biases (e.g., Naous et al. (2024); Kim and Lee (2024); Bajaj et al. (2024)) that can increase current health disparities mentioned in our discussion, there are additional concerns for using LLM-generated narratives to inform patients. Specifically, it needs to be made clear to patients that the narratives describe fictitious rather than real people. Knowing that the narratives do not describe real patient experiences could potentially make them less relevant and effective for patients than the literature suggests. Findings of the study by Boomstra et al. (2024) on narratives personalized with a template-based approach, for example, are based on participants thinking that they were reading narratives from real individuals. After being told that the narratives were personalized in an automated process, participants noted that it was important to them that the narratives were based on actual patient experiences and several felt a little betrayed. And also in work where LLM-generated narratives were just introduced as illustrating how somebody may navigate a situation and thus not explicitly as real (Bhattacharjee et al., 2025), it is not clear whether some participants thought the narratives were fully describing real experiences. While our lay participants evaluating narrative quality knew that narratives were generated by LLMs, our evaluation does not consider clinical effectiveness. Future work should also investigate how narratives can be clearly marked as fictitious and how clinically effective they then are.

Properly handling patient data. If improved versions of our narratives are ultimately generated based on real patient profiles in a clinical setting, care should be taken that patient data is handled properly. For example, by working only with anonymized and hence not identifiable profiles and passing these to locally running open-weight models as we did. Moreover, also for such low-risk use patients need to give consent.

Involving the target group. This work is a collaboration between NLP and clinical oncology researchers and thus draws on clinical knowledge such as the use of thresholds for clinical importance (Giesinger et al., 2020) when judging how accurately the narratives represent the PROM scores. However, so far no direct members of the target group (i.e., oncology patients and their healthcare providers) have been involved. Now that we have provided initial evidence for the potential of generating personalized narratives based on PROM profiles and we work toward using the narratives in a clinical setting, both oncology patients and their healthcare providers need to be involved if we want to make sure that the narratives promote their well-being.

Considering the environmental impact of language model choices. Our evaluation shows that `Llama-3.3-70B-Instruct` led to the overall best performance across the different criteria. However, besides performance, the costs of using a model should also be considered. This includes the direct costs that researchers or their institutes need

to pay for the resources, but also indirect externalities such as energy use, carbon emissions, and the consumption of clean water and minerals (e.g., Luccioni et al. (2023, 2024a,b)), whose effects tend to be disproportionally felt by already marginalized groups (Agyeman et al., 2003). While the smallest models we tested led to quite poor accuracy, it could be that fine-tuned small language models perform well. Such fine-tuning could potentially make use of the many interviews psychosocial oncology researchers conduct, thus also effectively reusing research data. This is worth investigating.

Data availability statement

All data and analysis code underlying this paper can be found in the online repository accompanying this paper (Albers et al., 2026). This includes the LLM-generated narratives and accuracy evaluations as well as the prompts and code used to generate them, data from the quality evaluation with human raters, and all analysis code that allows reproducing our analyses inside Docker containers as recommended by van de Schoot et al. (2021). Moreover, our public preregistration (He et al., 2026) not only provides detailed information on our analyses but also makes clear which parts of our analyses are exploratory versus confirmatory.

Acknowledgments

This work is funded by the IMPROVE project, an Innovative Health Initiative project granted by the European Commission under grant agreement ID: 101132847. NA is further funded by the project 'RISE: Relapse Intervention and Self-Management (Technology Assisted Self-Management: Preventing Relapse and Crisis by the Severe Mentally Ill Themselves)' with file number KICH1.GZ03.21.002 of the research program 'KIC - MISSIE Zorg in eigen leefomgeving 2021', which is (partly) financed by the Dutch Research Council (NWO). KdL and AT are funded by the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group for the project 'Activating cancer patients in research and care by effectively presenting PROs to patients using tailored narratives', grant number 006-2024.

References

Neil K. Aaronson, Sam Ahmedzai, Bengt Bergman, Monika Bullinger, Ann Cull, Nicole J. Duez, Antonio Filiberti, Henning Flechtner, Stewart B. Fleishman, Johanna C. J. M. de Haes, Stein Kaasa, Marianne Klee, David Osoba, Darius Razavi, Peter B. Rofe, Simon Schraub, Kommer Sneeuw, Marianne Sullivan, and Fumikazu Takeda. 1993. The european organization for research and treatment of cancer qlq-c30: A quality-of-life instrument for use in international clinical trials in oncology. *JNCI: Journal of the National Cancer Institute*, 85(5):365–376.

Julian Agyeman, Robert Doyle Bullard, and Bob Evans. 2003. *Just sustainabilities: Development in an unequal world*. MIT press.

Nele Albers, Linwei He, Anouk Teunissen, Kelly de Ligt, Nadine Bol, and Emiel Krahmer. 2026. Can we use LLMs to generate data-driven patient narratives? A multi-method exploration in the oncology domain – Data and analysis code.

Terrance L. Albrecht, Susan S. Eggly, Marci E.J. Gleason, Felicity W.K. Harper, Tanina S. Foster, Amy M. Peterson, Heather Orom, Louis A. Penner, and John C. Ruckdeschel. 2008. Influence of clinical communication on patients' decision making on participation in clinical trials. *Journal of Clinical Oncology*, 26(16):2666–2673. PMID: 18509178.

Sandra Andraszewicz, Benjamin Scheibehenne, Jörg Rieskamp, Raoul Grasman, Josine Verhagen, and Eric-Jan Wagenmakers. 2015. An introduction to bayesian hypothesis testing for management research. *Journal of Management*, 41(2):521–543.

Abeer Badawi, Md Tahmid Rahman Laskar, Elahe Rahimi, Sheri Grach, Lindsay Bertrand, Lames Danok, Frank Rudzicz, Jimmy Huang, and Elham Dolatabadi. 2026. Assessing the quality of mental health support in LLM responses through multi-attribute human evaluation. *Preprint*, arXiv:2601.18630.

Divij Bajaj, Yuanyuan Lei, Jonathan Tong, and Ruihong Huang. 2024. Evaluating gender bias of LLMs in making morality judgements. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15804–15818, Miami, Florida, USA. Association for Computational Linguistics.

Albert Bandura. 2004. Health promotion by social cognitive means. *Health Education & Behavior*, 31(2):143–164.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Hilary L Bekker, Anna E Winterbottom, Phyllis Butow, Amanda J Dillard, Deb Feldman-Stewart, Floyd J Fowler, Maria L Jibaja-Weiss, Victoria A Shaffer, and Robert J Volk. 2013. Do personal stories make patient decision aids more effective? A critical review of theory and evidence. *BMC Medical Informatics and Decision Making*, 13(Suppl 2):S9.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Rohit Bharti and Manmohan Sharma. 2026. Modeling emotional trajectories in storytelling LLMs: Toward emotionally coherent AI narratives for learning contexts. In *2026 Sixth International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies (ICAECT)*, pages 1–7.

Ananya Bhattacharjee, Joseph Jay Williams, Karrie Chou, Justice Tomlinson, Jonah Meyerhoff, Alex Mariakakis, and Rachel Kornfield. 2022. "I kind of bounce off it": Translating mental health principles into real life through story-based text messages. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2).

Ananya Bhattacharjee, Sarah Yi Xu, Pranav Rao, Yuchen Zeng, Jonah Meyerhoff, Syed Ishtiaque Ahmed, David C. Mohr, Michael Liut, Alex Mariakakis, Rachel Kornfield, and Joseph Jay Williams. 2025. Perfectly to a tee: Understanding user perceptions of personalized LLM-enhanced narrative interventions. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, DIS '25, page 1387–1416, New York, NY, USA. ACM.

Eva Boomstra, Saar Hommes, RD Vromans, S Van Der Burg, AM Schrijver, MWJM Wouters, IMC van der Ploeg, MW van de Kamp, EJ Krahmer, LV van de Poll-Franse, and KM de Ligt. 2024. "numbers call for action, personalized narratives provide support and recognition": A qualitative assessment of cancer patients' perspectives on patient-reported outcome measures (PROMs) feedback with narratives. *Journal of Cancer Survivorship*, 20(2):465–482.

Selmer Bringsjord and David A Ferruci. 1999. *Artificial Intelligence and Literary Creativity: Inside the mind of BRUTUS, a storytelling machine*. Psychology Press.

Mary Ann Burg, Gail Adorno, Ellen D. S. Lopez, Victoria Loerzel, Kevin Stein, Cara Wallace, and Dinghy Kristine B. Sharma. 2015. Current unmet needs of cancer survivors: Analysis of open-ended responses to the american cancer society study of cancer survivors ii. *Cancer*, 121(4):623–630.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28.

Melanie J Calvert, Samantha Cruz Rivera, Ameeta Retzer, Sarah E Hughes, Lisa Campbell, Barbara Molony-Oates, Olalekan Lee Aiyegbusi, Angela M Stover, Roger Wilson, Christel McMullan, Nicola E Anderson, Grace M Turner, Elin Haf Davies, Rav Verdi, Galina Velikova, Paul Kamudoni, Syed Muslim, Adrian Gheorghe, Daniel O'Connor, and 3 others. 2022. Patient reported outcome assessment must be inclusive and equitable. *Nature Medicine*, 28(6):1120–1124.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? Large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. ACM.

Richard A Chechile. 2020. *Bayesian Statistics for Experimental Scientists: A General Introduction Using Distribution-Free Methods*. MIT Press.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? Prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142.

Cohere. 2024. Command r: Retrieval-augmented generation at production scale. `https://cohere.com/blog/command-r`. Accessed on 26.09.2025.

ContactDoctor. 2024. Contactdoctor-bio-medical: A high-performance biomedical language model. https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B.

Kelly M de Ligt, Saar Hommes, Ruben D Vromans, Eva Boomstra, Lonneke V van de Poll, and Emiel J Krahmer. 2025. Improving the implementation of patient-reported outcome measure in clinical practice: Tackling current challenges with innovative digital communication technologies. *Journal of Medical Internet Research*, 27:e60777.

Jürgen Dietrich and André Hollstein. 2025. Performance and reproducibility of large language models in named entity recognition: Considerations for the use in controlled environments. *Drug Safety*, 48(3):287–303.

David C. Funder and Daniel J. Ozer. 2019. Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2):156—-168.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(1):65—-170.

C Gibbons, I Porter, DC Gonçalves-Bradley, S Stoilov, I Ricci-Cabello, E Tsangaris, J Gangannagaripalli, A Davey, EJ Gibbons, A Kotzeva, J Evans, PJ van der Wees, E Kontopantelis, J Greenhalgh, P Bower, J Alonso, and JM Valderas. 2021. Routine provision of feedback from patient-reported outcome measurements to healthcare providers and patients in clinical practice. *Cochrane Database of Systematic Reviews*, (10).

Johannes M. Giesinger, Fanny L.C. Loth, Neil K. Aaronson, Juan I. Arraras, Giovanni Caocci, Fabio Efficace, Mogens Groenvold, Marieke van Leeuwen, Morten Aa. Petersen, John Ramage, Krzysztof A. Tomaszewski, Teresa Young, and Bernhard Holzner. 2020. Thresholds for clinical importance were established to improve interpretation of the EORTC QLQ-C30 in clinical practice and research. *Journal of Clinical Epidemiology*, 118:1–8.

Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, 79(5):701–721.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 180 others. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Alexander Gurung and Mirella Lapata. 2025. Learning to reason for long-form story generation. In *Second Conference on Language Modeling*.

Linwei He, Nele Albers, and Emiel Krahmer. 2026. Evaluation of LLM-generated narratives based on patient-reported outcome measures.

S Ishaque, J Karnon, G Chen, R Nair, and AB Salter. 2019. A systematic review of randomised controlled trials evaluating the use of patient-reported outcome measures (PROMs). *Quality of Life Research*, 28(3):567–592.

Mete Ismayilzada, Claire E. Stevenson, and Lonneke van der Plas. 2025. Evaluating creative short story generation in humans and large language models. In *Proceedings of the 16th International Conference on Computational Creativity, Campinas, Brazil, June 23-27, 2025*, pages 145–155. Association for Computational Creativity (ACC).

Junsol Kim and Byungkyu Lee. 2024. AI-augmented surveys: Leveraging large language models and surveys for opinion prediction. *Preprint*, arXiv:2305.09620.

Matthew W. Kreuter, Melanie C. Green, Joseph N. Cappella, Michael D. Slater, Meg E. Wise, Doug Storey, Eddie M. Clark, Daniel J. O'Keefe, Deborah O. Erwin, Kathleen Holmes, Leslie J. Hinyard, Thomas Houston, and Sabra Woolley. 2007. Narrative communication in cancer prevention and control: A framework to guide research and application. *Annals of Behavioral Medicine*, 33(3):221–235.

Michael R. Krogsgaard, John Brodersen, Karl Bang Christensen, Volkert Siersma, Svend Kreiner, Jonas Jensen, Christian Fugl Hansen, and Jonathan D. Comins. 2021. What is a PROM and why do we need it? *Scandinavian Journal of Medicine & Science in Sports*, 31(5):967–971.

Cara H Lai, Lauren M Shapiro, Derek F Amanatullah, Loretta B Chou, Michael J Gardner, Serena S Hu, Marc R Safran, and Robin N Kamal. 2022. A framework to make PROMs relevant to patients: Qualitative study of communication preferences of PROMs. *Quality of Life Research*, 31(4):1093–1103.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CultureLLM: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

Lingyao Li, Deyi Li, Chen Chen, Renkai Ma, Runlong Yu, Mingquan Lin, Rui Yin, Lizhou Fan, Cathy Shyr, Siyuan Ma, Mei Liu, and Steven Bethard. 2026. Llm-as-a-judge in healthcare: A scoping analysis of applications, methods, and human alignment. *Preprint*, arXiv:2605.25273.

David Y. Liu, Aditya Joshi, and Paul Dawson. 2026. Narrative theory-driven llm methods for automatic story generation and understanding: A survey. *Preprint*, arXiv:2602.15851.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253):1–15.

Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024a. Power hungry processing: Watts driving the cost of AI deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 85–99, New York, NY, USA. ACM.

Sasha Luccioni, Bruna Trevelin, and Margaret Mitchell. 2024b. The environmental impacts of AI – Policy primer. In *Hugging Face Blog*. Accessed on 25.09.2025.

Richard McElreath. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.

Amy McQueen, Matthew W Kreuter, Bindu Kalesan, and Kassandra I Alcaraz. 2011. Understanding narrative effects: The impact of breast cancer survivor stories on message processing, attitudes, and beliefs among african american women. *Health Psychology*, 30(6):674–682.

Meta AI. 2024. Introducing meta llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/. Accessed on 26.09.2025.

Mistral AI. 2023. Mistral 7b. https://mistral.ai/news/announcing-mistral-7b. Accessed on 26.09.2025.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

Sheila T. Murphy, Lauren B. Frank, Joyee S. Chatterjee, and Lourdes Baezconde-Garbanati. 2013. Narrative versus nonnarrative: The role of identification, transportation, and emotion in reducing health disparities. *Journal of Communication*, 63(1):116–137.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? Measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. *Informatics*, 11(3).

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Sachin R. Pendse, Kate Niederhoffer, and Amit Sharma. 2019. Cross-cultural differences in the use of online mental health support forums. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

Farhat Tasnim Progga, Amal Khan, and Sabirat Rubya. 2024. Large language models and personalized storytelling for postpartum wellbeing. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '24, page 653–657, New York, NY, USA. ACM.

Farhat Tasnim Progga and Sabirat Rubya. 2026. Who tells the better story? Comparing human and AI-generated narratives on perinatal mental health topics. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. ACM.

Meghan Reading Turchioe, Lisa V Grossman, Annie C Myers, Dawon Baik, Parag Goyal, and Ruth M Masterson Creber. 2020. Visual analogies, not graphs, increase patients' comprehension of changes in their health status. *Journal of the American Medical Informatics Association*, 27(5):677–689.

Stefan Rennick-Egglestone, Kate Morgan, Joy Llewellyn-Beardsley, Amy Ramsay, Rose McGranahan, Steve Gillard, Ada Hui, Fiona Ng, Justine Schneider, Susie Booth, Vanessa Pinfold, Larry Davidson, Donna Franklin, Simon Bradstreet, Simone Arbour, and Mike Slade. 2019. Mental health recovery narratives and their impact on recipients: Systematic review and narrative synthesis. *The Canadian Journal of Psychiatry*, 64(10):669–679. PMID: 31046432.

Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–267.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Victoria A Shaffer, Elizabeth S Focella, Andrew Hathaway, Laura D Scherer, and Brian J Zikmund-Fisher. 2018. On the usefulness of narratives: An interdisciplinary review and theoretical model. *Annals of Behavioral Medicine*, 52(5):429–442.

Chantal Shaib, Venkata S Govindarajan, Joe Barrow, Jiuding Sun, Alexa Siu, Byron C Wallace, and Ani Nenkova. 2025. Standardizing the measurement of text diversity: A tool and comparative analysis. In *Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 36–46, Mumbai, India. Association for Computational Linguistics.

Amira Skeggs, Ashish Mehta, Valerie Yap, Seray B Ibrahim, Charla Rhodes, James J. Gross, Sean A. Munson, Predrag Klasnja, Amy Orben, and Petr Slovak. 2025. Micro-narratives: A scalable method for eliciting stories of people's lived experience. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. ACM.

Maria Teleki, Vedangi Bengali, Xiangjue Dong, Sai Tejas Janjur, Haoran Liu, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, and James Caverlee. 2025. A survey on LLMs for story generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13954–13966, Suzhou, China. Association for Computational Linguistics.

Stine Thestrup Hansen, Mette Kjerholt, Sarah Friis Christensen, John Brodersen, and Bibi Hølge-Hazelton. 2020. User experiences on implementation of patient reported outcome measures (PROMs) in a haematological outpatient clinic. *Journal of Patient-Reported Outcomes*, 4(1):87.

Rens van de Schoot, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, and Christopher Yau. 2021. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Esmee M. van der Willik, Caroline B. Terwee, Willem Jan W. Bos, Marc H. Hemmelder, Kitty J. Jager, Carmine Zoccali, Friedo W. Dekker, and Yvette Meuleman. 2021. Patient-reported outcome measures (PROMs): Making sense of individual PROM scores and changes in PROM scores over time. *Nephrology*, 26(5):391–399.

Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F Gronau, Martin Šmíra, Sacha Epskamp, Dora Matzke, Jeffrey N Rouder, and Richard D Morey. 2018. Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25(1):35–57.

Janine Westendorp, Olaf P. Geerse, Marije L. van der Lee, Jan W. Schoones, Milon H. M. van Vliet, Tamara Wit, Andrea W. M. Evers, and Liesbeth M. van Vliet. 2023. Harmful communication behaviors in cancer care: A systematic review of patients and family caregivers perspectives. *Psycho-Oncology*, 32(12):1827–1838.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. 2023. Faithful AI in medicine: A systematic review with large language models and beyond. *MedRxiv*.

Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre L. S. Filipowicz. 2023. More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, C&C '23, pages 368—370, New York, NY, USA. ACM.

Prompts

QL	PF	RF	EF	CF	SF	FA	NV	PA	DY	SL	AP	CO	DI	FI
75	100	100	58	100	100	11	0	0	0	67	0	33	0	0
100	100	100	75	100	100	11	0	17	0	33	0	0	0	0
100	93	100	100	100	100	0	0	0	0	0	0	0	0	0
83	100	100	100	100	100	0	0	17	0	0	0	0	0	0
67	93	100	92	100	100	22	0	17	0	0	0	0	0	0
50	100	67	42	83	83	0	0	0	0	67	33	0	0	0
75	60	33	42	50	50	67	33	67	67	67	33	0	33	33
100	100	100	100	100	100	0	0	0	0	0	0	0	0	0
75	87	100	67	83	100	22	0	17	0	67	0	0	0	0
33	20	0	58	0	0	100	67	83	67	100	100	0	100	33

Table A1: The ten real PROM profiles retrieved from anonymized data provided by the Netherlands Cancer Institute. Data is provided for the global health status (QL), physical functioning (PF), role functioning (RF), emotional functioning (EF), cognitive functioning (CF), social functioning (SF), fatigue (FA), nausea and vomiting (NV), pain (PA), dyspnea (DY), insomnia (SL), appetite loss (AP), constipation (CO), diarrhea (DI), and financial difficulties (FI).

Task description

You are a cancer patient. You are writing a short story reflecting on your daily experience. Your story should reflect your scores from the EORTC QLQ-C30 questionnaire.

Write a brief narrative describing your everyday experiences as a cancer patient, in a way that corresponds to your EORTC QLQ-C30 results.

As a cancer patient, you are asked to write a short story about your day-to-day life. Let your story reflect how you feel and function according to your EORTC QLQ-C30 scores.

Write a short, reflective story about your everyday life as a cancer patient that represents your EORTC QLQ-C30 profile.

Write a short personal story that captures what your daily life is like as a cancer patient. Make sure your story reflects your scores on the EORTC QLQ-C30 questionnaire.

Personal patient characteristics

You are a [age] year old [gender] dealing with [cancer]. You are [occupation] and you like [hobby].

Imagine you are a [age]-year-old [gender] who has been diagnosed with [cancer]. You are [occupation] and enjoy [hobby].

You are a [age]-year-old [gender] with [cancer], [occupation], and enjoying [hobby].

Take the perspective of a [age]-year-old [gender] diagnosed with [cancer]. You are [occupation] and like [hobby].

You are a [age]-year-old [gender] dealing with [cancer]. You are [occupation] and enjoy [hobby].

Explanation of EORTC-QLQ-C30 and patient's PROM scores

The EORTC QLQ-C30 is a questionnaire used to assess the quality of life of cancer patients. It consists of 14 subscales: Functioning scales include Physical, Role, Emotional, Cognitive, and Social functioning; Symptom scales include Fatigue, Nausea/Vomiting, Pain, Dyspnoea, Insomnia, Appetite Loss, Constipation, Diarrhea, and Financial Difficulties. An additional item Global Health Status/QoL indicates overall health and quality of life. Each subscale is scored from 0 to 100. For functioning scales and global health, a higher score indicates better functioning. For symptom scales, a higher score indicates more severe symptoms.

There are thresholds for clinical importance (TCI) for each subscale, which are specific score cutoffs that indicate when a symptom or functional problem is serious enough to limit daily life, require help or care, or cause worry. For Physical Functioning, Role Functioning, Emotional Functioning, Cognitive Functioning, and Social functioning, scores above the TCI suggest good functioning; for Fatigue, Nausea/Vomiting, Pain, Dyspnoea, Insomnia, Appetite Loss, Constipation, Diarrhea, and Financial Difficulties, scores above the TCI suggest the symptom may be concerning, and scores below the TCI suggest the symptoms are less worrisome. The distance between the score and the TCI indicates the severity or clinical relevance of a score.

Below are your scores for each subscale and whether they are better or worse than the TCI:

Global health status: [score], physical functioning: [score] ([better/worse] than TCI), role functioning: [score] ([better/worse] than TCI), emotional functioning: [score] ([better/worse] than TCI), cognitive functioning: [score] ([better/worse] than TCI), social functioning: [score] ([better/worse] than TCI), fatigue: [score] ([better/worse] than TCI), nausea and vomiting: [score] ([better/worse] than TCI), pain: [score] ([better/worse] than TCI), dyspnoea: [score] ([better/worse] than TCI), insomnia: [score] ([better/worse] than TCI), appetite loss: [score] ([better/worse] than TCI), constipation: [score] ([better/worse] than TCI), diarrhea: [score] ([better/worse] than TCI), financial difficulties: [score] ([better/worse] than TCI). Remember to reflect the meaning of the scores in your story.

Your story should mention all 14 scales in the questionnaire. Do not mention the scores. Make sure to correctly reflect if a functioning aspect and symptom is worrisome or not.

Stylistic guidelines

Write in a way that brings the reader into your world, use vivid and specific descriptions. Let the reader see through your eyes - share your thoughts, feelings, and personal reactions. Include emotions, describe how your experiences make you feel.

Figure A1: Narrative generation prompt template. For each prompt component we created five paraphrases, which we show for the task description and personal patient characteristics. The paraphrases for the other prompt components can be found in the online repository accompanying this paper.

Task description

Your task is to determine, for each aspect of the EORTC QLQ-C30 questionnaire, whether the patient's experience was described in the story and whether the story suggests that their score is above or below the threshold of clinical importance.

Narrative

This is the story: '[narrative]'

Explanation of the EORTC QLQ-C30

The EORTC QLQ-C30 is a questionnaire used to assess the quality of life of cancer patients. It consists of 14 subscales: Functioning scales include Physical, Role, Emotional, Cognitive, and Social functioning; Symptom scales include Fatigue, Nausea/Vomiting, Pain, Dyspnoea, Insomnia, Appetite Loss, Constipation, Diarrhea, and Financial Difficulties. For functioning scales, a higher score indicates better functioning. For symptom scales, a higher score indicates more severe symptoms.

There are thresholds for clinical importance (TCI) for each subscale, which are specific score cutoffs that indicate when a symptom or functional problem is serious enough to limit daily life, require help or care, or cause worry. For Physical Functioning, Role Functioning, Emotional Functioning, Cognitive Functioning, and Social functioning, scores above the TCI suggest good functioning; for Fatigue, Nausea/Vomiting, Pain, Dyspnoea, Insomnia, Appetite Loss, Constipation, Diarrhea, and Financial Difficulties, scores above the TCI suggest the symptom may be concerning, and scores below the TCI suggest the symptoms are less worrisome.

The TCI values are:

physical functioning: 83, role functioning: 58, emotional functioning: 71, cognitive functioning: 75, social functioning: 58, fatigue: 39, nausea and vomiting: 8, pain: 25, dyspnea: 17, insomnia: 50, appetite loss: 50, constipation: 50, diarrhea: 17, financial difficulties: 17.

Output instructions

Please return your answer as a list using only one word per subscale: “absent” or either "above" or "below" the TCI. "absent" thereby means that the subscale is not described in the story.
Do not include any explanation, reasoning, or additional text. Only output the 14 lines in the exact format below.

Format (example):

Physical functioning: above

Role functioning: below

Emotional functioning: above

Cognitive functioning: absent

Social functioning: above

Fatigue: below

Nausea and vomiting: below

Pain: above

Dyspnoea: below

Insomnia: below

Appetite loss: below

Constipation: above

Diarrhea: above

Financial difficulties: below

Make sure to include one judgment ("above", "below", or "absent") for each of the 14 subscales, do not skip any, even if the narrative gives limited information.

Figure A2: Narrative evaluation prompt template.

Agreement between LLM judges

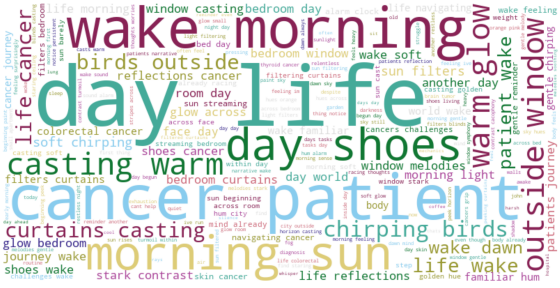
Subscale	Fleiss κ
Physical functioning	0.75
Role functioning	0.64
Emotional functioning	0.24
Cognitive functioning	0.59
Social functioning	0.66
Fatigue	0.78
Nausea and vomiting	0.75
Pain	0.74
Dyspnea	0.79
Insomnia	0.82
Appetite loss	0.78
Constipation	0.82
Diarrhea	0.80
Financial difficulties	0.87
Average	**0.72**

Table B1: Fleiss κ between the three LLM judges for each subscale of the EORTC QLQ-C30.

What do the generated narratives look like?

Model	Words M (SD)	Sentences M (SD)	Words per sentence M (SD)
`gpt-oss-20b medium`	771.76 (234.45)	43.53 (14.29)	17.19 (7.94)
`gpt-oss-20b low`	794.75 (175.51)	44.66 (13.35)	**17.29** (7.96)
`s1.1-32B`	**904.65** (252.09)	**72.61** (24.08)	<u>11.13</u> (6.01)
`DeepSeek-R1-Distill-Qwen-32B`	558.59 (171.32)	36.52 (11.98)	14.48 (6.17)
`Mistral-7B-Instruct-v0.3`	801.76 (542.59)	45.62 (40.40)	17.16 (7.51)
`Command-r-v01`	520.18 (91.90)	35.38 (8.29)	12.62 (7.52)
`Llama-3.3-70B-Instruct`	593.28 (73.54)	36.57 (6.60)	15.52 (7.26)
`Llama-3.1-8B-Instruct`	680.31 (89.59)	43.76 (7.53)	15.16 (5.70)
`Llama-3.2-3B-Instruct`	664.79 (248.00)	39.85 (16.78)	16.25 (5.97)
`Bio-Medical-Llama-3-8B`	<u>432.82</u> (144.12)	<u>28.70</u> (10.75)	14.43 (6.33)

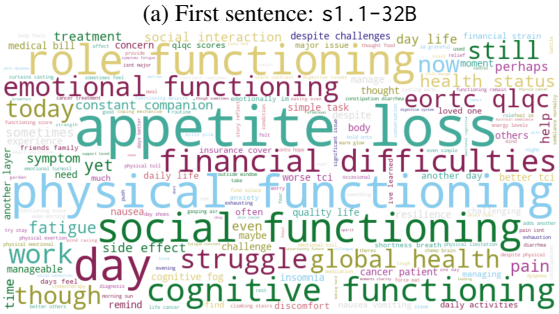
Table C1: Mean number of words, sentences, and words per sentence for the narratives generated by the ten models. The **highest** and <u>lowest</u> numbers are highlighted.



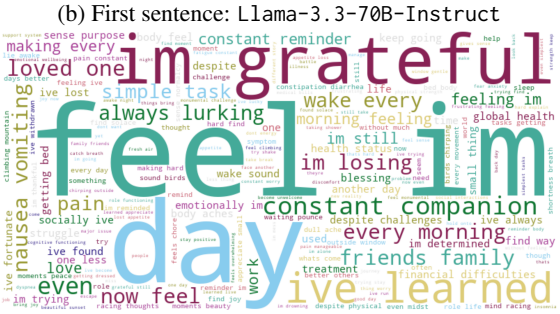
(a) First sentence: `s1.1-32B`



(b) First sentence: `Llama-3.3-70B-Instruct`



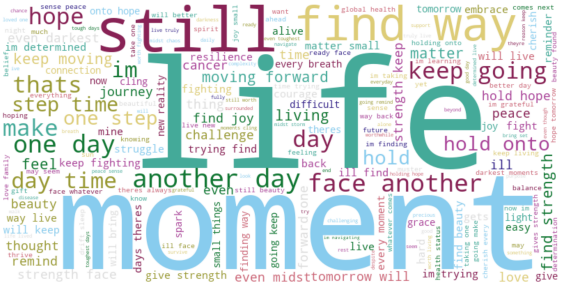
(c) Full narrative: `s1.1-32B`



(d) Full narrative: `Llama-3.3-70B-Instruct`



(e) Last sentence: `s1.1-32B`



(f) Last sentence: `Llama-3.3-70B-Instruct`

Figure C1: Wordclouds of first sentences, full narratives, and last sentences generated by `s1.1-32B` and `Llama-3.3-70B-Instruct`. Wordclouds for the other models can be found in the online repository accompanying this paper.

Model	1-gram	2-gram	3-gram	4-gram
gpt-oss-20b medium	0.000038	0.000525	0.003248	0.011979
gpt-oss-20b low	0.000035	0.000502	0.003164	0.011635
s1.1-32B	**0.000254**	**0.001315**	0.004379	0.013088
DeepSeek-R1-Distill-Qwen-32B	0.000203	0.000977	0.004354	0.014412
Mistral-7B-Instruct-v0.3	0.000032	0.000440	0.002716	0.009761
Command-r-v01	0.000053	0.000735	**0.004693**	**0.017232**
Llama-3.3-70B-Instruct	0.000040	0.000480	0.002935	0.010525
Llama-3.1-8B-Instruct	0.000038	0.000489	0.002906	0.010339
Llama-3.2-3B-Instruct	0.000039	0.000463	0.002806	0.009911
Bio-Medical-Llama-3-8B	0.000065	0.000857	0.004593	0.015277

Table C2: N-gram diversity computed as the ratio of unique n-grams to the total number of n-grams for the narratives generated by the ten models. The highest diversity scores are highlighted in bold.

Model and narrative
gpt-oss-20b medium. The first thing I notice when I step into the museum is the hum of the crowd, the faint scent of old varnish, and the way light catches on a bronze statue. It's a place where time slows, and for a moment I feel as if I'm back in my twenties, walking barefoot on marble floors, unburdened by the weight of a brain tumor that sits quietly in my skull. I walk up the spiral staircase, the steps smooth under my feet. My legs feel strong, my body moves without the heaviness that sometimes creeps in after a long day. I can reach for the next exhibit without hesitation, and I can keep up with the tour guide's lively narration. That is my physical functioning—steady and reliable. When I think about my daily chores, I can still manage the laundry, the dishes, the occasional grocery run, all without feeling drained. That's my role functioning, and it's a comfort to know I still hold a place in the rhythm of my home. The museum's quiet corners, where I can sit and read a placard, bring me a sense of peace that I rarely find elsewhere. My mind is sharp; I can remember the names of artists, the dates of movements, the details of a painting's brushwork. My cognitive functioning is a bright spot in my day. I still feel connected to the world, even as my body reminds me of the fragility of my health. I laugh with a friend at a quirky sculpture, and I feel the warmth of human connection. My social functioning is strong—I'm not shy about asking questions, sharing my thoughts, or simply enjoying the company of others. The museum is a bridge between my inner world and the outside, a place where I can feel both independent and part of a larger community. When I leave the museum, I can still walk home without feeling the sharp sting of pain or the tightening in my chest. I have no nausea, no vomiting, no aches that make me pause. My breathing is effortless, my lungs feel open and clear. I can lie down after a day of walking and feel a sense of relief, knowing that I haven't been overwhelmed by any physical discomfort. But there are moments that tug at me harder than the rest. The night after a long day of exploring, I lie awake, the ceiling above my head a shifting constellation of thoughts. Sleep eludes me, and the quiet of the house feels louder than the museum's chatter. Insomnia is a quiet, persistent companion that reminds me of the restlessness inside my mind. Emotionally, I feel a tremor of unease that I can't quite shake. The knowledge that I'm living with a tumor, that my brain is a place of both wonder and worry, sometimes makes me feel like I'm walking on a tightrope. There are days when I feel a weight in my chest, a sense of loss for the future that I had imagined. It's a reminder that my emotional functioning is not as robust as my physical and cognitive strengths.

Table C3: Narratives generated by different models for the first real PROM scores and a patient described as a 67-year-old woman with a brain tumor who is retired and likes visiting museums.

Model and narrative
Yet, even in those moments of discomfort, I find solace in the simple act of breathing. I count my breaths, I listen to the hum of the refrigerator, and I remind myself that my appetite remains intact; I still enjoy a good meal, the flavors of fresh bread, the aroma of coffee. I am not battling constipation or diarrhea—my digestive system is a quiet, reliable ally. I have no financial worries; the care I receive is covered, and I can focus on the present rather than the future. When I look back at the day, I see a mosaic of strengths and challenges. My global health feels steady, a gentle reminder that I am still living, still experiencing the beauty of the world. I can walk through a gallery, feel the texture of a canvas, and know that my brain tumor does not define me. It is a part of my story, but it is not the only chapter. In the quiet moments, I hold onto the hope that each day brings new light, new colors, and new stories to add to the canvas of my life.
s1.1-32B. **A Day in My World** I wake up to the soft chirping of birds outside my window, their melodies a gentle reminder that another day has begun. As I lie in bed, the first thing I notice is the persistent ache behind my eyes, a constant companion since the diagnosis. But today feels different. Maybe it's the sunlight filtering through the curtains, casting a warm glow across my room. Or perhaps it's the sense of peace that settles over me as I start my day. **Physical Functioning:** Despite the brain tumor, my body feels surprisingly strong. I get out of bed without assistance, moving through my morning routine with ease. The physical aspects of my life aren't hindered; I can still dress myself, prepare meals, and go about my daily activities. The doctors must be doing their job well. The treatment hasn't left me bedridden or weakened, which is a relief. I suppose that's why my physical functioning isn't an issue—I can still do what I need to do without feeling overly fatigued. **Role Functioning:** Being retired, my roles are different now. I don't have work obligations or responsibilities that demand my time and energy. Instead, my days are filled with hobbies and personal interests. The only "role" I have is being a wife and grandmother, but even those don't require the same level of engagement as before. So, role functioning isn't affected. I can still participate in family events, attend gatherings, and maintain my social connections. Though sometimes, the cognitive fog makes it challenging to follow conversations, especially during long discussions. **Emotional Functioning:** Here's where things get tricky. Emotionally, I struggle. The brain tumor affects my mood swings and anxiety levels. Some days, I feel like I'm on an emotional rollercoaster. One moment, I'm cheerful and optimistic, thinking about future museum visits; the next, I'm overwhelmed by sadness or fear. These fluctuations are exhausting. I've learned to accept them as part of my journey, but they do take a toll. Coping mechanisms like meditation and talking to my support group help manage these ups and downs, but there are days when I just want to curl up and cry. That's okay too—allowing myself to feel vulnerable is important. **Cognitive Functioning:** Another area impacted by the tumor is my cognition. Concentration is hit-or-miss. Some days, I can read a book or engage in a complex conversation; other days, simple tasks become daunting. The medical team reassures me that this is normal, given the location of the tumor. They emphasize the importance of rest and mental stimulation. I try to balance activities that challenge my mind with those that relax it. Puzzles and crosswords help keep my brain active, while listening to music or spending time in nature provides respite. **Social Functioning:** Maintaining relationships is crucial for me. Despite some days feeling drained, I prioritize staying connected with friends and family. We have regular phone calls, video chats, and occasional meet-ups. Social interactions give me purpose and joy. Even when I don't feel up to much, their support and love mean the world. The social aspect of my life remains strong, thanks to their unwavering presence. **Fatigue:** Interestingly, fatigue isn't a significant issue. Maybe because the treatments I'm on don't cause extreme tiredness. Or perhaps my body is adapting. Either way, I don't spend my days battling exhaustion. This is a blessing, as it allows me to remain engaged in activities I enjoy. However, I still listen to my body and rest when needed, understanding that pacing myself is key.

Table C3: (continued)

Model and narrative
Nausea/Vomiting: Thankfully, neither nausea nor vomiting has been a problem. The medication regimen must be effective in managing these side effects. I haven't experienced any episodes, which is a huge relief. Eating becomes a pleasure rather than a chore, and I savor each meal, appreciating the flavors and textures. **Pain:** Pain isn't a concern for me right now. The tumor's location might not press on sensitive areas, or the treatment is keeping it at bay. I don't take painkillers regularly, which is a positive sign. When I do have mild headaches, over-the-counter remedies suffice. It's a small victory, as pain can drastically affect one's quality of life. **Dyspnea:** Breathing comes easily. No shortness of breath here. The treatments aren't affecting my respiratory system, so I can move around freely without gasping for air. Simple tasks like walking to the kitchen or going upstairs don't leave me winded. That's a relief, as some cancer treatments can impact lung function. **Insomnia:** Ah, sleep. This is where I face challenges. Insomnia plagues me more than I'd like to admit. Lying awake at night, thoughts racing, worrying about treatment outcomes, wondering about the future. Sometimes, I toss and turn for hours, unable to find comfort. The mind becomes a battleground, replaying scenarios and fears. Medication helps occasionally, but I prefer natural remedies like chamomile tea and lavender oil. Sleepless nights are tough, but I try not to let them consume me. The next day, I remind myself that restful nights will come again. **Appetite Loss:** Fortunately, my appetite remains intact. I still enjoy meals and snacks, savoring each bite. The treatments haven't dulled my taste buds, so food is a source of pleasure. I cook for myself, experimenting with new recipes, finding solace in the process. Eating well is essential, so I pay attention to nutrition, ensuring I get the necessary nutrients to support my body through treatment. **Constipation:** This is a minor concern. While not severe, I do experience some constipation. Not uncommon, given the medications. I drink plenty of water, eat fiber-rich foods, and stay active. It's manageable, but something I monitor. Diarrhea hasn't been an issue, thankfully. **Financial Difficulties:** Thankfully, our finances are stable. Insurance covers most expenses, and we don't face major financial strain. That's one less worry. Some people deal with the stress of medical bills, but we've been fortunate. Still, I keep track of costs and look for ways to minimize unnecessary expenditures. **Global Health Status:** Considering everything, my overall health is decent. The tumor is a constant presence, but the treatments seem to be working. I have good days and bad days, but the majority lean toward the good. The support from loved ones, combined with my own resilience, keeps me going. There are moments of doubt, but I choose to focus on gratitude for each day. As I sit by the window, sipping my morning coffee, I reflect on how far I've come. Cancer is a journey filled with uncertainties, but I face it with hope. Each day is a gift, and I cherish the ability to enjoy simple pleasures—a walk in the park, a good book, a visit to a museum. These moments of joy sustain me through the tougher times. While some aspects of my life are affected, many remain vibrant. I am determined to live fully, despite the challenges. After all, every day is a masterpiece waiting to be painted. And I intend to color it with courage, grace, and a dash of whimsy. **End of Story**
`DeepSeek-R1-Distill-Qwen-32B`. **A Day at the Museum** The alarm buzzed softly at 7 AM, but I knew I wouldn't be able to sleep anymore. The night had been restless, a common companion these days. I tossed and turned, my mind racing with thoughts that refused to let me rest. Insomnia had become a frequent visitor, and tonight was no exception. I groaned as I sat up, the sheets tangled around my legs. The clock on my nightstand read 3:15 AM. Another sleepless night. I decided to get up and make some tea. The warmth of the cup in my hands was soothing, a small comfort in the quiet of the early morning. The house was still, save for the gentle hum of the heating system. I stood by the window, watching the city awaken, the first signs of dawn painting the sky in hues of orange and pink. It was a peaceful moment, a brief respite from the chaos in my mind.

Table C3: (continued)

Model and narrative
After a quick shower, I dressed in my favorite jeans and a cozy sweater, ready for the day. Today, I had planned a visit to the art museum. It was a place of solace, a sanctuary where I could lose myself in the beauty of another world. The museum was bustling when I arrived. The familiar buzz of conversation and the soft murmur of guides filled the air. I walked through the entrance, the cool air conditioning a welcome contrast to the warmth outside. The first exhibit I encountered was a collection of impressionist paintings. I stood before Monet's water lilies, the vibrant colors and soft brushstrokes drawing me in. For a moment, the weight of my thoughts lifted, and I felt a sense of peace. As I wandered through the galleries, I noticed a group of schoolchildren gathered around a sculpture. Their laughter and curiosity were infectious, and I found myself smiling. Social interactions came easily to me, a testament to my strong social functioning. I even struck up a brief conversation with a fellow visitor, sharing my appreciation for a particularly striking piece. The museum's café was a quiet spot where I could rest and reflect. I ordered a cappuccino and sat by the window, watching the world go by. The constipation I occasionally experienced was a minor nuisance, but it didn't deter me from enjoying my day. I sipped my coffee, savoring the rich flavor, and let my mind drift. Lunch was a simple sandwich, eaten in the quiet of the library section. The intellectual stimulation of the books and the quiet environment were a balm to my soul. My cognitive functioning remained sharp, allowing me to engage deeply with the ideas presented in the texts. As the day wore on, I found myself drawn to the modern art section. The abstract pieces challenged my perceptions, and I felt a sense of exhilaration as I tried to interpret their meanings. The emotional lows I often felt were a shadow in the background, but for now, the art lifted my spirits. The walk home was a pleasant one, the sun casting a golden glow on the streets. I felt a sense of accomplishment, having spent the day doing something I loved. The physical exertion of walking was manageable, and my body felt strong and capable. As I unlocked my front door, I felt a mix of emotions. The day had been a reminder of the resilience I still possessed, despite the challenges I faced. The emotional struggles were ever-present, but moments of joy and peace were still within reach. I knew the night would come again, with its restless hours and racing thoughts. But for now, I was content. The museum had been a refuge, a place where I could find solace and beauty, a reminder that life, in all its complexity, was still worth living. `Command-r-v01`. As I stand before the Monet exhibit, his vibrant waterlilies floating on the surface of the serene river, I feel a sense of peace wash over me. The colors, the light, the very essence of the painting speaks to my soul. I am at my favorite museum, and cancer, mere weeks after my diagnosis, is but a distant whisper in the chaos of my mind. I've always found sanctuary here among the masters, each brushstroke a reminder of the beauty and richness of life. It's not that I don't have bad days. Some mornings I wake up and the mere act of getting dressed feels like climbing Mount Everest. On those days, my physical functioning seems to have taken a hike. Getting through the fatigue, the cloud that wraps around my mind, feels like an Everest of its own. But today is not one of those days. Today, I feel alive, and my spirit soars as I wander through the galleries. Emotionally, things are a different story. Cancer has this insidious way of creeping up on me, a constant, quiet reminder that sits heavy on my chest. There are days I find myself overwhelmed by the simplest of tasks, the future I had envisioned for myself crumbles at the edges. It's like looking at a distorted funhouse version of myself, fragments of who I am scattered and warped. Those around me seem to notice a change, too. They tiptoe around me, unsure of what to say, their eyes full of pity. I want to scream that it's okay, that I'm still me, but the walls of my emotional functioning are crumble.

Table C3: (continued)

Model and narrative
But here, in this museum, surrounded by art, none of that matters. As I walk through the halls, my mind finds clarity. My cognitive functioning is sharp, and my soul feels alive. I connect the dots between the masterworks, losing myself in the evolution of an artist's journey. My steps are steady, and my physical being is strong. I take in each exhibit, each piece of art a reminder that the world is so much bigger than my diagnosis. Financial difficulties are the last thing on my mind; I am lucky—each step is a choice, a privilege some might not understand. The only distressing physical symptom is the insomnia that leaves me with a constant fatigue, a fog that clouds my mind. It's a heavy veil I can't seem to shake, and it worries me at times. What if this is the new normal? I haven't yet accepted this as part of my life, this weary, tired existence. Appetite comes and goes in waves; today it's a mere thought, a ghost of its former self. But the taste of the museum café's lemon tart lingers delightfully on my tongue, a welcome surprise after weeks of turning my nose up at the very mention of food. I find myself slowing down near the impressionist paintings, their blurry edges and vivid colors speaking to the chaos within me. It's as if the artists reached into my mind and translated the turmoil onto the canvas. I feel seen, understood. My social interactions are far from troubled; the museum goes about its day undeterred, and rightly so. I blend into the background as I observe the world around me, a world that continues turning. The only real pain is the ache in my heart, a deep, aching throb that never quite goes away. It's a constant reminder of my mortality, a drumbeat to the rhythm of my existence. But today, I refuse to let it win. Today, I stand tall, a warrior in front of a Monet, and I feel alive. As I leave, the museum's serene atmosphere lingers, calming the turmoil within me. I take a deep breath, grateful for the good days, knowing the fight within me is far from over. I may have a brain tumor, but it doesn't define me. Today, I define me. And today, I am a woman of strength, beauty, and fierce determination. The scores, the numbers, they mean nothing here. They don't see the true picture, the whole canvas. Because on days like this, I win. And that's all that matters.
`Llama-3.3-70B-Instruct`. As I sit here, surrounded by the familiar comforts of my home, I'm reminded of the simple joys that bring me peace. My physical strength remains, and I'm grateful for that. I can still take long walks, visit the museums I love, and tend to my garden without feeling overwhelmed. My role in life, though it has changed since retirement, still feels fulfilling. I manage my days with purpose, whether it's volunteering or spending time with loved ones. My cognitive abilities, thankfully, remain sharp, allowing me to enjoy puzzles, read extensively, and engage in meaningful conversations. Socially, I've been blessed with a supportive network of friends and family, who make social interactions a pleasure, not a chore. However, despite these blessings, there's a shadow that follows me - the emotional weight of my condition. It's as if a filter has been placed over my usually optimistic outlook, tinting everything with a hue of worry and sadness. There are days when the thought of what's to come feels overwhelming, and the fear of the unknown grips me tightly. It's a constant companion, one that I wish I could shake off, but it clings to me, affecting my mood and overall sense of well-being. Sleep, or the lack thereof, has become a significant challenge. Insomnia has made nights long and restless, leaving me tired and irritable in the mornings. It's a vicious cycle, where the exhaustion exacerbates my emotional fragility, making every challenge feel like a monumental task. Yet, in the midst of these struggles, there are aspects of my life that remain untouched by my illness. I've been fortunate enough not to experience the debilitating symptoms that I've heard others speak of - the nausea, the pain, the difficulty breathing. My appetite remains healthy, and I find joy in cooking and sharing meals with loved ones. The practical aspects of my life, like my finances, are also under control, which is a relief.

Table C3: (continued)

Model and narrative
But what strikes me most, as I navigate this journey, is the contrast between the parts of my life that are thriving and those that are suffering. My global sense of health and quality of life is a mixed bag. On one hand, I feel blessed with the strength and support that allows me to live a relatively normal life. On the other, the emotional and psychological impacts of my condition weigh heavily on me, affecting my overall sense of well-being and happiness. As I look back, I realize that this journey has taught me the value of each facet of my life. The physical, role, emotional, cognitive, and social aspects - each plays a vital role in my quality of life. And the symptoms, or lack thereof, each has its impact. It's a delicate balance, one that I strive to maintain, even on the toughest of days. And so, I hold on to hope, and the beauty of life, as seen through the lens of my experiences, becomes even more precious and worthy of appreciation.
`Llama-3.1-8B-Instruct`. I woke up to the warm sunlight streaming through the window, a gentle reminder that another day had begun. My husband, John, was already in the kitchen, brewing a pot of coffee. I smiled, feeling grateful for the comfort of home and the love that surrounded me. My physical health had been a concern lately, but thankfully, I'd been managing well. I could still move about with ease, take long walks in the park, and even tackle the stairs without getting winded. The physical exertion I'd been doing was helping, and I felt stronger each day. As I sipped my coffee, I thought about the art museum we'd visited last week. I'd always loved art, and the museum's latest exhibit had been a treat. The vibrant colors, the masterpieces, the history – it all came alive for me. My social life was thriving, thanks to John's encouragement and our shared love of culture. We'd met friends there, and the afternoon had been filled with laughter and conversation. I felt grateful for the social connections that kept me grounded. However, as I began to get ready for the day, I couldn't shake off the fatigue that had been creeping up on me. It wasn't overwhelming, but it was there, a gentle nudge that told me to slow down. I'd been taking naps in the afternoons, and that seemed to help. I wasn't too concerned about it, but I knew I had to keep an eye on it. Breakfast was a struggle, as it often was. My appetite had been a bit off, and I'd been experiencing some constipation. It wasn't severe, but it was a nuisance. I'd been eating smaller meals, trying to manage my portions, and drinking plenty of water. I wasn't too worried, but I knew I had to stay on top of it. As I sat down to eat, I felt a pang of sadness. My emotional well-being had been a bit of a challenge lately. The diagnosis of my brain tumor had been a shock, and it still felt surreal at times. I'd been feeling down, and it was hard to shake off the worry and fear that lingered in the back of my mind. It was a heavy burden to carry, but I was trying to stay positive. My mind was foggy at times, and I'd been experiencing some memory lapses. It was a concern, but I'd been working with my doctors to manage it. Cognitive function was a worry, but I was determined to stay sharp. I'd been doing puzzles, reading, and engaging in activities that challenged my mind. The morning passed quickly, and before I knew it, John and I were on our way to the museum. We spent the day surrounded by masterpieces, and I felt my spirits lift. The art was a balm to my soul, and I felt more connected to the world around me. As the sun began to set, we headed home, tired but content. My financial situation was stable, thanks to John's support and our careful planning. We'd been through ups and downs, but we'd always managed to stay afloat. As I drifted off to sleep that night, I felt a sense of peace. My global health was a concern, but I was managing. I was grateful for the love and support that surrounded me, and I knew that I'd get through this.
`Llama-3.2-3B-Instruct`. I wake up every morning, feeling a sense of relief wash over me as I realize I'm still able to get out of bed. My body, though, tells a different story. The fatigue that's been my constant companion for months has taken its toll, leaving me feeling drained and exhausted. I have to force myself to get dressed and make it to the kitchen for a cup of coffee. The morning light streaming through the window highlights the cluttered countertops, a reminder of the chaos that's become my life.

Table C3: (continued)

Model and narrative
As I sip my coffee, I'm grateful for the role I've managed to maintain in my social life. Despite the challenges, I still find myself attending museum exhibitions, lost in the world of art and history. The cognitive function that's served me well in my career as a retired librarian is still sharp, and I find myself lost in the intricate details of the museum's collection. However, emotions are a different story. The emotional toll of my diagnosis has left me feeling anxious and irritable at times. The emotional ups and downs are unpredictable, and I've learned to be kind to myself when I'm struggling. I've also noticed that my emotional well-being is closely tied to my social connections. I've been fortunate to have a strong support system, and I cherish the time I spend with loved ones. As I make my way to the museum, I'm grateful for the physical health that's allowed me to keep up with my favorite activities. My body may not be as energetic as it used to be, but I've learned to pace myself and take breaks when needed. I've also had to adapt to some changes in my daily routine, like taking medication to manage my symptoms. One of the things that's been a challenge is dealing with the nausea and vomiting that comes with my treatment. It's a constant companion, one that I've learned to live with. I've found ways to manage it, like taking breaks during the day and staying hydrated. Pain is another symptom I've learned to live with. It's not something I can see or touch, but it's very real. I've found ways to cope with it, like meditation and deep breathing exercises. Dyspnea is a concern, but I've been fortunate to not experience it as much as others. I've learned to pace myself and take breaks when needed. Insomnia has been a challenge, but I've found ways to manage it, like establishing a consistent sleep schedule and creating a relaxing bedtime routine. Appetite loss has been a problem, but I've learned to eat smaller, more frequent meals to keep my energy up. Constipation has been a concern, but I've found ways to manage it, like staying hydrated and taking medication as prescribed. Diarrhea has been a problem, but I've learned to stay hydrated and take breaks when needed. Financial difficulties haven't been a concern for me, but I've had to be mindful of my expenses and make adjustments to my budget. As I stand in front of a beautiful painting, I'm struck by the beauty of the world around me. The museum is a place of solace, a place where I can escape the challenges of my daily life. I feel grateful for the health and well-being that's allowed me to experience this world in all its beauty. Despite the challenges, I'm determined to make the most of every day. I know that life is precious, and I'm grateful for every moment I have. As I leave the museum, I feel a sense of peace, knowing that I've taken care of myself and made the most of my day.
`Bio-Medical-Llama-3-8B`. I'll do my best to write a story reflecting on my daily experience as a cancer patient, taking into account the scores provided. Here is my story: As I wake up each morning, I try to push aside the fatigue that often comes with my condition. It's not always easy, but I've learned to pace myself and take regular breaks to rest. I'm grateful that I still have the energy to do the things I love, like visiting museums. However, I do struggle with insomnia from time to time. Lying awake at night, thinking about all the things I need to do the next day, can be frustrating. But I've found that writing down my thoughts and making a to-do list helps me fall asleep faster. I'm also dealing with constipation, which can be uncomfortable and affect my mood. But I've learned to eat more fiber-rich foods and drink plenty of water to help prevent it. Despite these challenges, I try to stay positive and focus on the things I can do. I still enjoy spending time with friends and family, and I'm grateful for their support. My story reflects the scores provided for each subscale of the EORTC QLQ-C30 questionnaire.

Table C3: (continued)

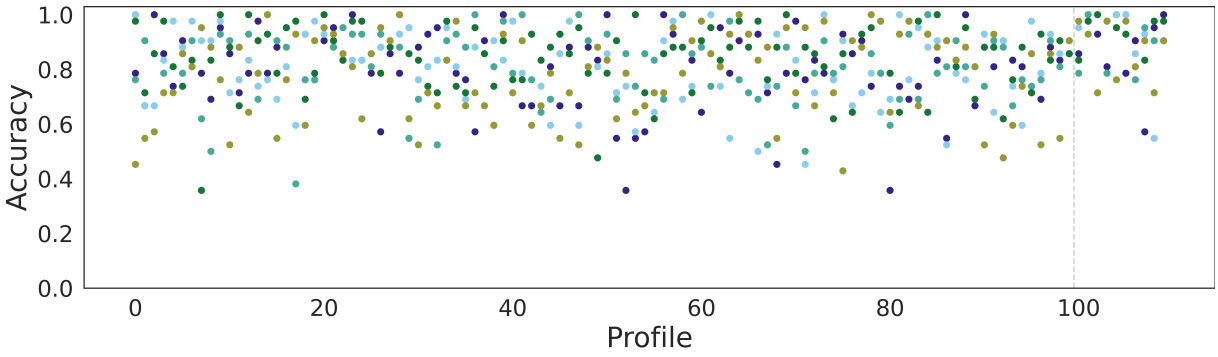
How accurate are the generated narratives?

Model	Accuracy M (SD), Range
gpt-oss-20b medium	0.746 (0.08), 0.61 – 0.82
gpt-oss-20b low	0.758 (0.08), 0.64 – 0.89
s1.1-32B	0.816 (0.02), 0.79 – 0.85
DeepSeek-R1-Distill-Qwen-32B	0.574 (0.03), 0.54 – 0.61
Mistral-7B-Instruct-v0.3	0.558 (0.03), 0.50 – 0.60
Command-r-v01	0.433 (0.03), 0.38 – 0.48
Llama-3.3-70B-Instruct	0.764 (0.07), 0.65 – 0.84
Llama-3.1-8B-Instruct	0.471 (0.04), 0.39 – 0.52
Llama-3.2-3B-Instruct	0.433 (0.04), 0.38 – 0.49
Bio-Medical-Llama-3-8B	0.464 (0.04), 0.40 – 0.54

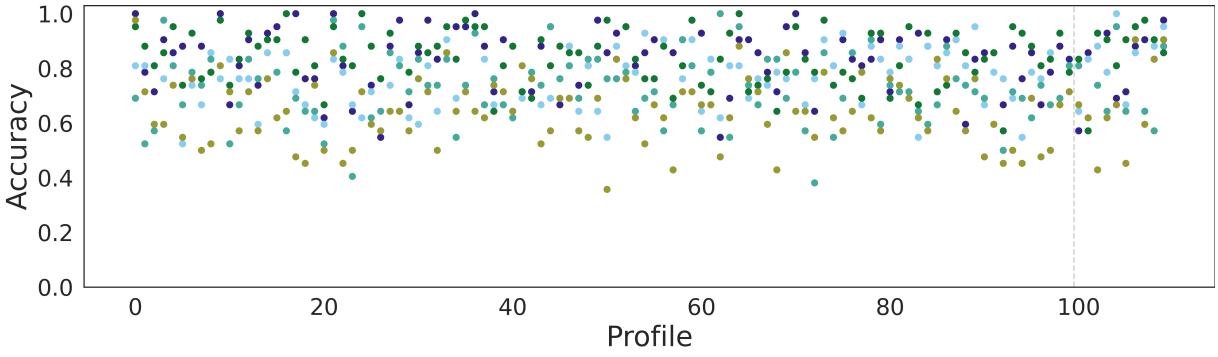
Table D1: Mean accuracy across the five paraphrases per model.

Model	Estimate	95% HDI	Prob(>0)	Interpretation
gpt-oss-20b medium	0.04	[0.02, 0.06]	>0.99995	Virtually certain
gpt-oss-20b low	0.06	[0.04, 0.07]	>0.99995	Virtually certain
DeepSeek-R1-Distill-Qwen-32B	0.25	[0.23, 0.26]	>0.99995	Virtually certain
Mistral-7B-Instruct-v0.3	0.26	[0.24, 0.28]	>0.99995	Virtually certain
Command-r-v01	0.39	[0.37, 0.40]	>0.99995	Virtually certain
Llama-3.3-70B-Instruct	0.05	[0.04, 0.07]	>0.99995	Virtually certain
Llama-3.1-8B-Instruct	0.35	[0.33, 0.37]	>0.99995	Virtually certain
Llama-3.2-3B-Instruct	0.39	[0.37, 0.40]	>0.99995	Virtually certain
Bio-Medical-Llama-3-8B	0.36	[0.34, 0.38]	>0.99995	Virtually certain

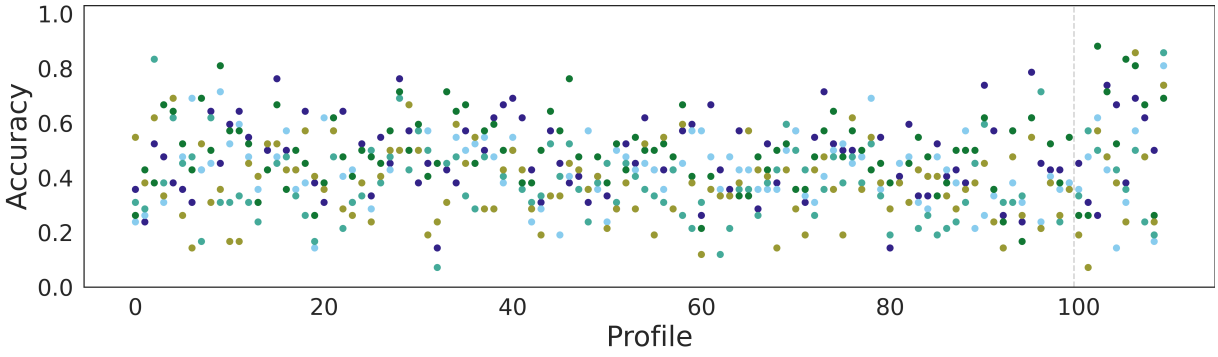
Table D2: Results of Bayesian analyses comparing s1.1-32B as the model with the overall highest accuracy to all other models in pairwise comparisons. We provide the estimate of the fixed effect for the accuracy of s1.1-32B being higher than that of the respective other model. The interpretations of the posterior probabilities that the fixed effect is greater than zero are based on Chechile (2020).



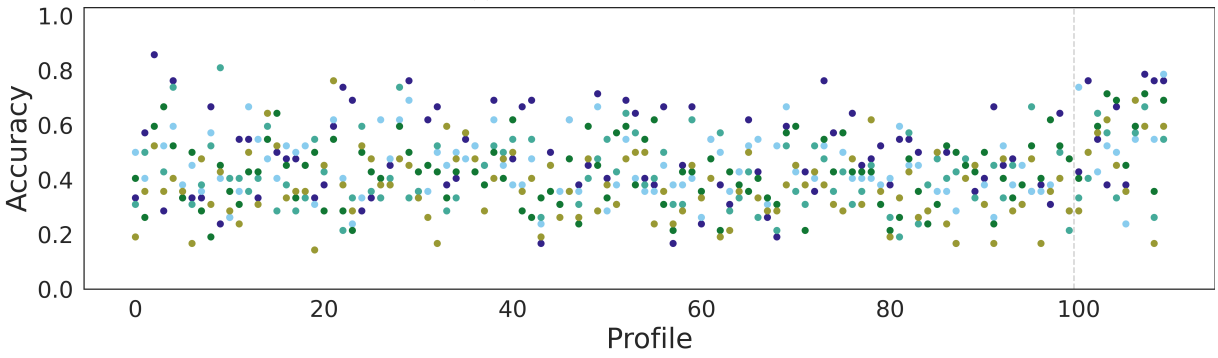
(a) s1.1-32B



(b) Llama-3.3-70B-Instruct



(c) Llama-3.2-3B-Instruct



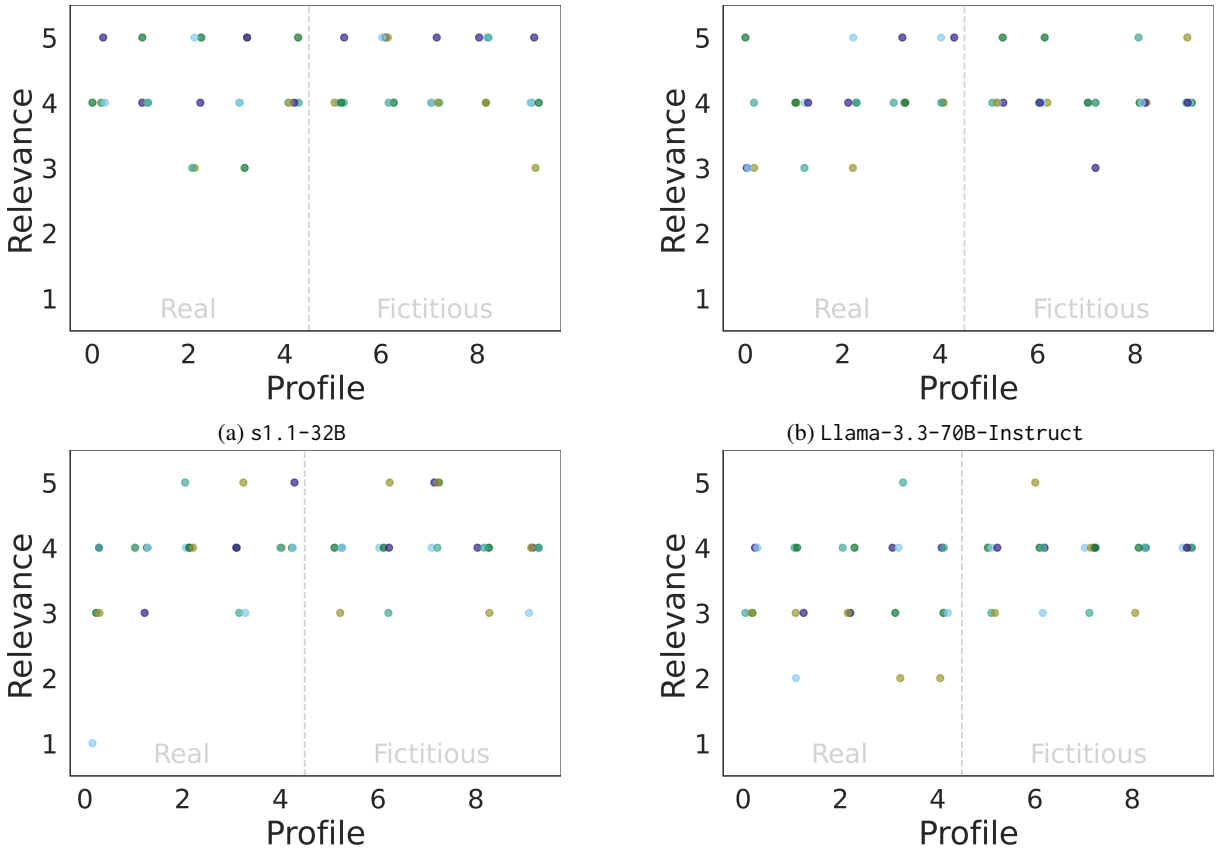
(d) Command-r-v01

Figure D1: Accuracy for the 5 prompt paraphrases for each of the 110 PROM profiles for the two models with the overall highest mean accuracy and those with the overall lowest mean accuracy. The different colors denote the 5 prompt paraphrases. The vertical lines separate the 100 fictitious from the 10 real PROM profiles.

Subscale	Accuracy M (SD), Range
Physical functioning	0.73 (0.04), 0.66 – 0.78
Role functioning	0.70 (0.03), 0.65 – 0.75
Emotional functioning	0.71 (0.04), 0.64 – 0.75
Cognitive functioning	0.74 (0.04), 0.68 – 0.78
Social functioning	**0.76** (0.02), 0.73 – 0.79
Fatigue	0.61 (0.03), 0.57 – 0.65
Nausea and vomiting	0.50 (0.07), 0.39 – 0.57
Pain	0.58 (0.03), 0.52 – 0.61
Dyspnea	0.48 (0.07), 0.38 – 0.55
Insomnia	0.54 (0.03), 0.48 – 0.57
Appetite loss	0.66 (0.03), 0.60 – 0.69
Constipation	0.42 (0.08), 0.30 – 0.51
Diarrhea	0.35 (0.08), 0.23 – 0.44
Financial difficulties	0.64 (0.05), 0.56 – 0.69

Table D3: Mean accuracy across the 5 prompt paraphrases for each of the 14 subscales of the EORTC QLQ-C30. The scores of all ten models are combined. The **highest** and lowest mean scores are highlighted.

What is the perceived quality of the narratives?



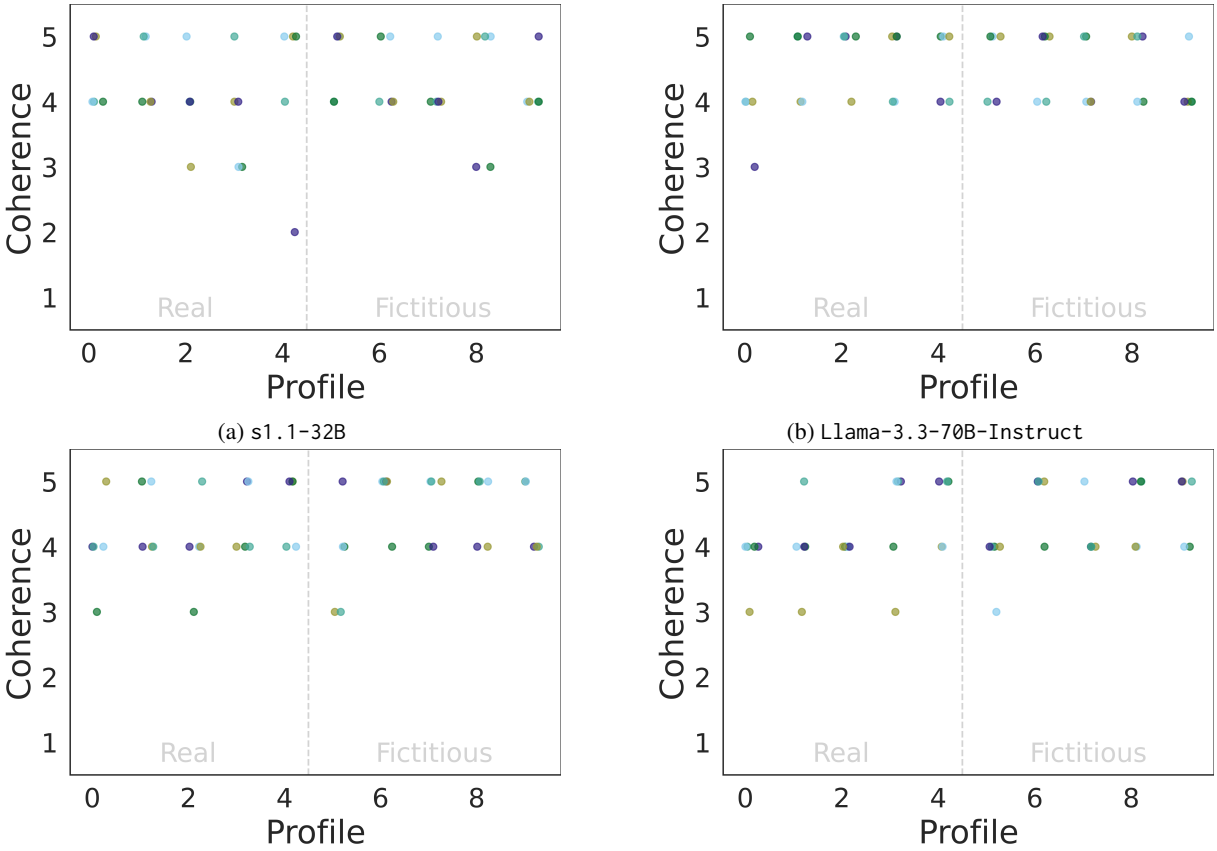
(a) `s1.1-32B` (b) `Llama-3.3-70B-Instruct`

(c) `Llama-3.2-3B-Instruct` (d) `Command-r-v01`

Figure E1: Relevance for the 5 prompt paraphrases for each of the 10 PROM profiles used in the quality evaluation for the two models with the overall highest mean accuracy and those with the overall lowest mean accuracy. For this figure, the scores by the three human evaluators of a narrative were averaged and rounded to the nearest integer. The different colors denote the 5 prompt paraphrases and the vertical line separates the real and the fictitious PROM profiles.

Quality measure	**Model 1** WAIC (SE)	**Model 2** WAIC (SE)
Relevance	3999.9 (44.7)	**3969.3** (44.8)
Coherence	3491.5 (54.6)	**3478.7** (54.7)
Empathy	4206.5 (43.0)	**4195.8** (43.6)
Engagement	4430.2 (38.2)	**4418.5** (38.4)

Table E1: For each of the four quality measures, we show the WAIC of two Bayesian models. The first model contains a general intercept and random intercepts for the paraphrases and PROM profiles, and the second model additionally contains random intercepts for the ten LLMs. A lower WAIC indicates a better model. The lower WAIC score is shown in bold.



(a) s1.1-32B

(b) Llama-3.3-70B-Instruct

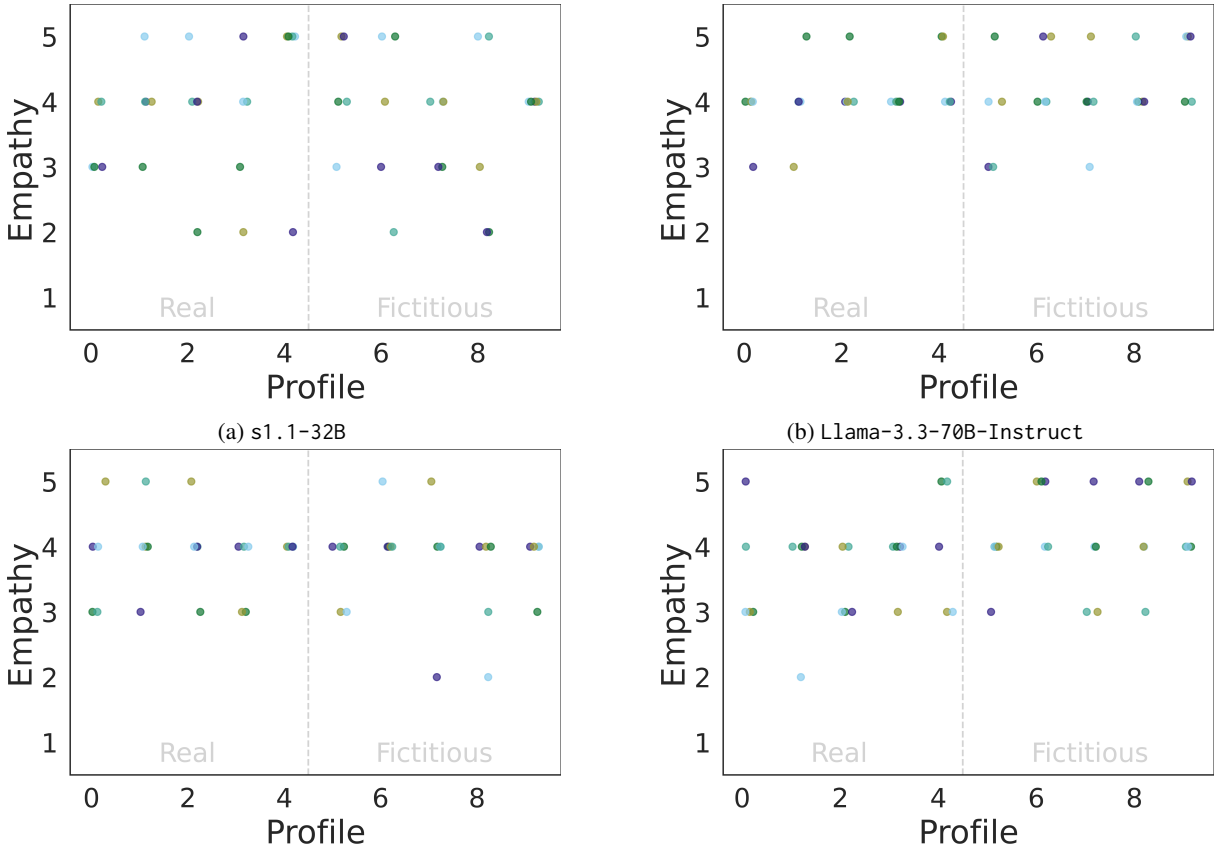
(c) Llama-3.2-3B-Instruct

(d) Command-r-v01

Figure E2: Coherence for the 5 prompt paraphrases for each of the 10 PROM profiles used in the quality evaluation for the two models with the overall highest mean accuracy and those with the overall lowest mean accuracy. For this figure, the scores by the three human evaluators of a narrative were averaged and rounded to the nearest integer. The different colors denote the 5 prompt paraphrases and the vertical line separates the real and the fictitious PROM profiles.

Model	**Estimate**	95% HDI	Prob(>0)	**Interpretation**
gpt-oss-20b medium	0.40	[0.15, 0.67]	0.9982	Very strong bet
gpt-oss-20b low	0.44	[0.19, 0.70]	0.9995	Nearing certainty
DeepSeek-R1	0.43	[0.18, 0.71]	>0.99995	Virtually certain
Mistral-7B	0.32	[0.09, 0.58]	0.9982	Very strong bet
Command-r-v01	0.68	[0.42, 0.97]	>0.99995	Virtually certain
Llama-3.3-70B-Instruct	0.19	[-0.05, 0.45]	0.9390	A promising but risky bet
Llama-3.1-8B-Instruct	0.40	[0.14, 0.67]	0.9993	Very strong bet
Llama-3.2-3B-Instruct	0.47	[0.20, 0.77]	0.9995	Nearing certainty
Bio-Medical-Llama-3-8B	0.51	[0.27, 0.79]	>0.99995	Virtually certain

Table E2: Results of Bayesian analyses comparing s1.1-32B as the model with the overall highest mean relevance score to all other models in pairwise comparisons. We provide the estimate of the fixed effect for the relevance scores of s1.1-32B being higher than those of the respective other model. The interpretations of the posterior probabilities that the fixed effect is greater than zero are based on Chechile (2020).



(a) s1.1-32B

(b) Llama-3.3-70B-Instruct

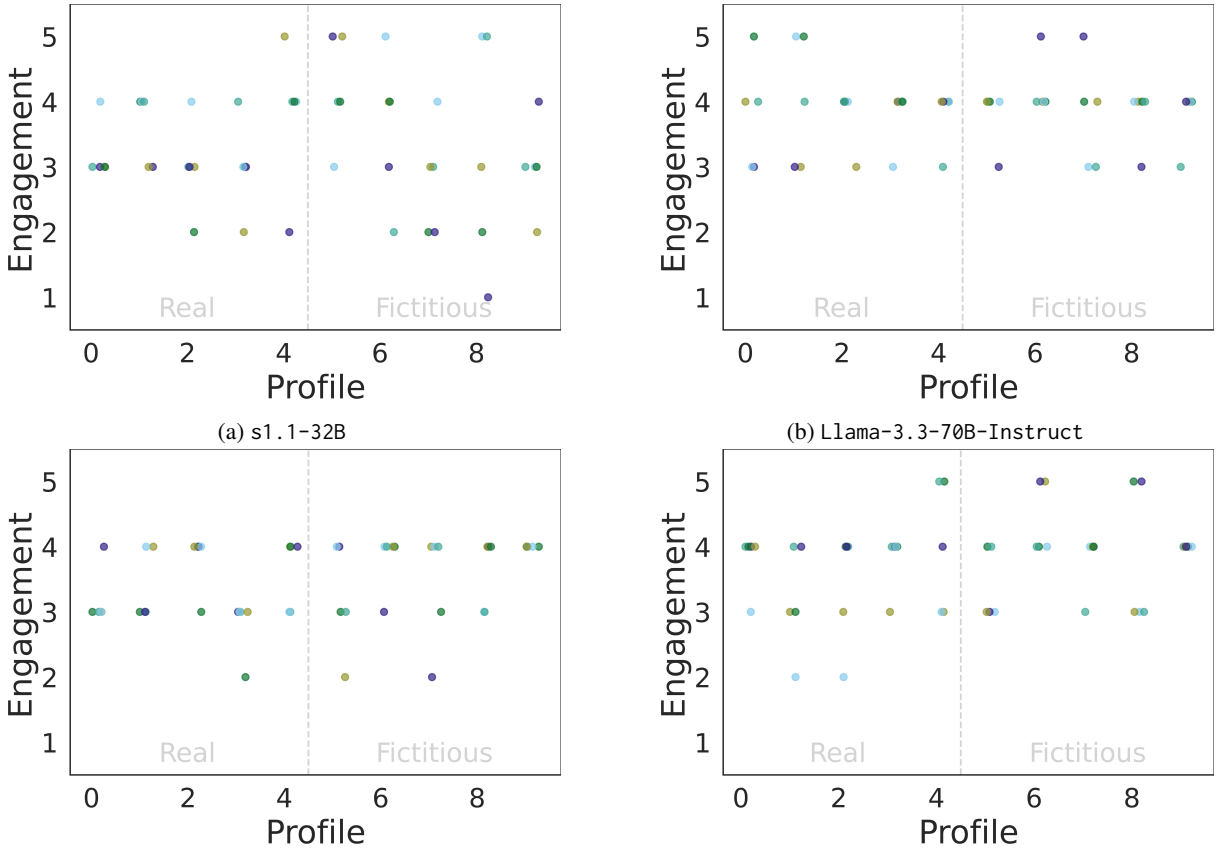
(c) Llama-3.2-3B-Instruct

(d) Command-r-v01

Figure E3: Empathy for the 5 prompt paraphrases for each of the 10 PROM profiles used in the quality evaluation for the two models with the overall highest mean accuracy and those with the overall lowest mean accuracy. For this figure, the scores by the three human evaluators of a narrative were averaged and rounded to the nearest integer. The different colors denote the 5 prompt paraphrases and the vertical line separates the real and the fictitious PROM profiles.

Model	**Estimate**	95% HDI	Prob(>0)	**Interpretation**
gpt-oss-20b medium	0.27	[0.00, 0.57]	0.9737	Good bet - too good to disregard
gpt-oss-20b low	0.46	[0.18, 0.79]	0.9995	Nearing certainty
s1.1-32B	0.21	[-0.05, 0.50]	0.9367	A promising but risky bet
DeepSeek-R1	0.15	[-0.13, 0.46]	0.8460	Only a casual bet
Mistral-7B	0.45	[0.18, 0.78]	0.9997	Nearing certainty
Command-r-v01	0.14	[-0.13, 0.43]	0.8345	Only a casual bet
Llama-3.1-8B-Instruct	0.13	[-0.16, 0.46]	0.8007	Only a casual bet
Llama-3.2-3B-Instruct	0.14	[-0.15, 0.46]	0.8258	Only a casual bet
Bio-Medical-Llama-3-8B	0.19	[-0.09, 0.49]	0.8987	Only a casual bet

Table E3: Results of Bayesian analyses comparing Llama-3.3-70B-Instruct as the model with the overall highest mean coherence score to all other models in pairwise comparisons. We provide the estimate of the fixed effect for the coherence scores of Llama-3.3-70B-Instruct being higher than those of the respective other model. The interpretations of the posterior probabilities that the fixed effect is greater than zero are based on Chechile (2020).



(a) s1.1-32B

(b) Llama-3.3-70B-Instruct

(c) Llama-3.2-3B-Instruct

(d) Command-r-v01

Figure E4: Engagement for the 5 prompt paraphrases for each of the 10 PROM profiles used in the quality evaluation for the two models with the overall highest mean accuracy and those with the overall lowest mean accuracy. For this figure, the scores by the three human evaluators of a narrative were averaged and rounded to the nearest integer. The different colors denote the 5 prompt paraphrases and the vertical line separates the real and the fictitious PROM profiles.

Model	**Estimate**	95% HDI	Prob(>0)	**Interpretation**
gpt-oss-20b medium	0.30	[0.04, 0.58]	0.9885	Good bet - too good to disregard
gpt-oss-20b low	0.25	[-0.05, 0.57]	0.9465	A promising but risky bet
s1.1-32B	0.34	[0.07, 0.63]	0.9940	Strong bet - irresponsible to avoid
DeepSeek-R1	0.32	[0.04, 0.62]	0.9872	Good bet - too good to disregard
Mistral-7B	0.29	[0.03, 0.58]	0.9867	Good bet - too good to disregard
Command-r-v01	0.12	[-0.13, 0.39]	0.8180	Only a casual bet
Llama-3.1-8B-Instruct	0.04	[-0.24, 0.32]	0.6007	Not worth betting on
Llama-3.2-3B-Instruct	0.37	[0.09, 0.67]	0.9963	Very strong bet
Bio-Medical-Llama-3-8B	0.53	[0.26, 0.83]	>0.99995	Virtually certain

Table E4: Results of Bayesian analyses comparing Llama-3.3-70B-Instruct as the model with the overall highest mean empathy score to all other models in pairwise comparisons. We provide the estimate of the fixed effect for the empathy scores of Llama-3.3-70B-Instruct being higher than those of the respective other model. The interpretations of the posterior probabilities that the fixed effect is greater than zero are based on Chechile (2020).

Model	Estimate	95% HDI	Prob(>0)	Interpretation
gpt-oss-20b medium	0.30	[0.06, 0.55]	0.9928	Strong bet - irresponsible to avoid
gpt-oss-20b low	0.39	[0.14, 0.66]	0.9990	Very strong bet
s1.1-32B	0.33	[0.08, 0.62]	0.9957	Very strong bet
DeepSeek-R1	0.20	[-0.05, 0.45]	0.9442	A promising but risky bet
Mistral-7B	0.24	[0.02, 0.47]	0.9848	Good bet - too good to disregard
Command-r-v01	0.05	[-0.19, 0.30]	0.6503	Not worth betting on
Llama-3.1-8B-Instruct	0.02	[-0.23, 0.28]	0.5557	Not worth betting on
Llama-3.2-3B-Instruct	0.25	[0.01, 0.51]	0.9818	Good bet - too good to disregard
Bio-Medical-Llama-3-8B	0.31	[0.07, 0.57]	0.9935	Strong bet - irresponsible to avoid

Table E5: Results of Bayesian analyses comparing Llama-3.3-70B-Instruct as the model with the overall highest mean engagement score to all other models in pairwise comparisons. We provide the estimate of the fixed effect for the engagement scores of Llama-3.3-70B-Instruct being higher than those of the respective other model. The interpretations of the posterior probabilities that the fixed effect is greater than zero are based on Chechile (2020).

Quality measure	Estimate	95% HDI	P(>0)	Interpretation
Relevance	-0.18	[-0.41, 0.06]	0.06	A promising but risky bet against
Coherence	-0.03	[-0.19, 0.12]	0.34	Not worth betting against
Empathy	-0.13	[-0.30, 0.04]	0.05	A promising but risky bet against
Engagement	-0.10	[-0.25, 0.05]	0.08	A promising but risky bet against

Table E6: For each of the four quality measures, we show the fixed effect of whether the real PROM profiles receive higher quality ratings than the fictitious profiles. The interpretations of the posterior probabilities that the fixed effect is greater than zero are based on Chechile (2020) and their extension to values below 0.5 by Andraszewicz et al. (2015).

Relation between accuracy and quality

Model and narrative
`gpt-oss-20b medium`. I wake up before the sun, the old wooden floorboards of my shop creaking under my feet. The world feels heavy, as if the very air is thick with a weight I can't lift. I try to get out of bed, but my body feels like a broken machine—every movement is a reminder that my physical functioning is far below what it once was. The simple act of stretching my arms feels like a climb up a steep hill, and I'm left with a sense of helplessness that settles over me like a blanket. The day starts with a rush of tasks: the register, the inventory of yarns and beads, the orders that need to be packed. My role functioning is shattered; I can't fully engage with the responsibilities that used to bring me joy. I feel like a ghost in my own shop, watching the customers with a detached eye, unable to offer the warm smiles and enthusiastic guidance that were my trademark. The emptiness in my role is a quiet ache that gnaws at me. Emotionally, I'm a storm of worry and frustration. The cancer's shadow looms over every thought, and I find myself sinking into a deep, relentless fatigue that doesn't lift even with a cup of tea or a short walk. I'm exhausted all the time, and the fatigue is so intense that I can't focus on the next stitch or the next pattern. It's a constant, oppressive presence that makes even the simplest tasks feel monumental. My mind, once sharp and quick, now struggles to keep up. I'm having trouble concentrating on the designs I love, and the cognitive fog is a relentless companion. I find myself forgetting the names of colors or the steps in a complex pattern, and this loss of mental clarity is a source of quiet despair. Socially, I've retreated. Friends and family used to drop by, sharing laughter and stories, but now I feel like an inconvenience. I'm too tired to answer calls or meet for coffee, and the isolation is a quiet, aching void. I miss the shared laughter, the simple connection that once felt effortless. The physical symptoms are relentless. Nausea and vomiting come on unexpectedly, a sudden wave that knocks me off balance. The pain is sharp and throbbing, a constant reminder of the battle I'm fighting. Breathing feels labored; dyspnea is a constant companion, making it hard to breathe in the quiet of the night. Insomnia has become my nightly routine—my mind racing, my body refusing to rest, and I wake up in a cold sweat, feeling more drained than before I went to bed. Appetite has vanished. I can't bring myself to eat, and when I do, the food feels like a burden rather than nourishment. The loss of appetite is a silent thief, stealing the energy I need to survive. Diarrhea follows, a painful, urgent need that disrupts my day and adds another layer of discomfort. Yet, oddly, I haven't had any constipation—an unexpected relief in the midst of all these troubles. Financial difficulties are a looming cloud. Running a small business has always been a passion, but the costs of treatment, medications, and the time I can no longer spend at the shop are draining my savings. The worry about bills, rent, and the future is a constant pressure, a weight that presses against my chest whenever I think about the next month. The global picture of my health feels like a dim, flickering light. I know I'm not where I want to be, and the sense of low overall health is a heavy, unshakable feeling that lingers, even when I try to focus on the bright colors of my yarns or the delicate sparkle of beads. I still love my crafts. The idea of creating something beautiful from raw materials gives me a sliver of hope, a small spark that refuses to be extinguished. But each day is a battle, a struggle to keep the flame alive amid the darkness of my illness. I keep telling myself that I'll get through this, that the joy of crafting will bring me back, but the reality of my functioning—my physical, emotional, cognitive, and social struggles—reminds me that the road ahead is long and fraught with challenges.
`s1.1-32B`. **A Day in My Skin Cancer Journey**

Table E7: The two generated narratives with the overall highest scores for quality and accuracy. The narrative by `gpt-oss-20b medium` (i.e., `gpt-oss-20b` with a medium reasoning level) received the maximum score for relevance, coherence, empathy, and accuracy, and a mean score of 4.33 out of 5 for engagement. The narrative by `s1.1-32B` received the maximum score for relevance, coherence, and accuracy, a mean score of 4.67 for empathy, and a mean score of 4.33 for engagement. Quality and accuracy are weighed equally to determine the narratives with the overall highest scores.

Model and narrative
I wake up before dawn, my body already protesting the effort. The sheets are tangled around me, a sweaty mess from another night of tossing and turning. Insomnia clutches at me relentlessly, making sleep a distant memory. Each morning, I struggle to peel myself out of bed, my limbs heavy with exhaustion. Fatigue is my constant companion, sapping my energy and leaving me breathless. Even simple tasks—brushing my teeth, getting dressed—feel Herculean. My husband helps me into the bathroom, where I face the mirror. The sight of my altered appearance still shocks me. Scars mar my once-smooth skin, a stark reminder of the battle raging inside me. The oncologist's words echo in my mind: "We caught it early, but we need to be thorough." The treatment plan is aggressive, designed to eradicate every trace of cancer. But with each passing day, I feel more like a shell of my former self. Physical functioning is a challenge. Walking from the bedroom to the kitchen feels like a marathon. My joints ache, and my muscles weaken. The once-familiar routine of running my small craft store seems like a lifetime away. Now, even sitting upright for extended periods is difficult. The physical toll is relentless, and I often wonder if I'll ever regain my strength. Role functioning has plummeted. As a business owner, I used to thrive on the chaos of inventory management, customer interactions, and creative problem-solving. Now, those responsibilities weigh heavily on me. Decision-making is foggy; concentration eludes me. Cognitive functioning is a battlefield. Simple tasks require intense focus, and sometimes I find myself staring at a spreadsheet, unable to process the numbers. The chemo brain is real, and it's frustrating. I rely on my husband's support more than I ever thought possible. Emotionally, I oscillate between despair and resilience. Some days, I'm consumed by fear and uncertainty. Will the treatments work? What if the cancer spreads? These thoughts spiral, feeding on each other. Other days, I find solace in crafting, the tactile nature of working with materials grounding me. Yet, the emotional rollercoaster takes its toll. I cry easily, laughing just as readily. Mood swings are unpredictable, leaving me feeling raw and exposed. Social functioning has suffered. Before, I was the life of the party, hosting craft nights and networking events. Now, I avoid social situations whenever possible. The idea of putting on a brave face while battling nausea and pain is exhausting. Friends check in, but I often decline their invitations. I don't want to burden them with my struggles. Isolation becomes a coping mechanism, though it deepens my loneliness. Symptoms dominate my existence. Nausea is a constant, making me wary of food. Even the smell of certain scents triggers a wave of queasiness. I've lost weight, despite trying to maintain my appetite. Eating feels like a battle. The chemotherapy-induced diarrhea adds another layer of discomfort. Each trip to the bathroom is a source of anxiety, wondering when the next episode will strike. Meanwhile, constipation is a rare blessing, a respite from the chaos. Pain is a persistent companion. The surgical sites throb, and neuropathy causes tingling sensations in my extremities. Every day, I take medications to manage the discomfort, but they only dull the edge. The pain reminds me constantly that I am not invincible. Dyspnea, too, is a frequent visitor. Shortness of breath makes even the simplest activities challenging. Climbing stairs leaves me gasping, clutching the railing for support. Each breath feels like an effort, a reminder of my fragile state. Financial difficulties loom large. Medical bills stack up, and the uncertainty of my prognosis weighs on our finances. Insurance covers much, but the copays and out-of-pocket expenses add up. We've had to dip into savings, a source of stress. The thought of being unable to pay for treatment haunts me. Money worries compound the emotional strain, creating a vicious cycle. Yet, amidst the turmoil, moments of beauty still exist. Crafting provides a sanctuary. When my hands are busy with beads or fabric, my mind quiets. There's a sense of control in creating something tangible, a counterpoint to the chaos within. These moments are fleeting but vital, reminding me that I'm more than my diagnosis.

Table E7: (continued)

Model and narrative
As night falls, I lie in bed, the day's challenges receding into the background. Despite the physical and emotional toll, there's a part of me that remains unbroken. Resilience is a muscle I didn't know I had, one that strengthens with each passing day. I may be grappling with cancer, but I refuse to let it define me entirely. This journey is a crucible, forging a new version of myself—one who can endure, adapt, and perhaps, emerge wiser and more compassionate. **Reflection:** This narrative encapsulates the experiences of a 47-year-old woman with skin cancer based on her EORTC QLQ-C30 profile. Her global health status is low (33.3), indicating significant overall impairment. Physical functioning (20) and role functioning (0) are severely compromised, reflecting her struggle with daily activities and business responsibilities. Emotional (58) and cognitive (0) functioning are also significantly affected, with mood swings and cognitive fog. Social functioning (0) is impacted by isolation. Symptoms like fatigue (100), nausea/vomiting (67), pain (83), dyspnea (67), insomnia (100), appetite loss (100), and diarrhea (100) are major concerns, while constipation (0) is less problematic. Financial difficulties (33) add stress. The story conveys her daily battles, emotional turmoil, and moments of hope, painting a vivid picture of life with cancer.

Table E7: (continued)

Model and narrative
`Command-r-v01`. The days merge into one another, an endless blur since my diagnosis. The fight against cancer is relentless, a brutal battle that consumes my every waking moment. The fatigue is overwhelming, a weight on my shoulders, making even the simplest tasks feel like climbing mountains. I wake up with the dull ache of pain already present, a constant companion. The night before, sleep was elusive, another symptom that torments me, leaving me exhausted before the day even begins. I feel the strain of it all—the treatments, the appointments, the endless pills and their side effects. My body doesn't belong to me anymore, each reaction and sensation a mystery, a new symptom to decipher and discuss with the doctors. The financial strain is a constant hum in the background, a dull ache that amplifies my stress. Work is a distant memory—I'm consumed by the role change, a forced retirement, and the loss of my once-active role in society. My days now revolve around clinics and hospitals, a new routine that I never imagined. I struggle to find joy and meaning, a victim of my circumstances, feeling depressed and angry at the cards I've been dealt. It's a daily battle just to find the energy to step outside, to connect with others. The world has become a blurry, grim place, the vibrant hues drained away, leaving a muted palette in its wake. But there is good—the support of family and friends lifts me up when I feel at my lowest. I cling to their encouragement, the understanding of those who know my struggle. My appetite, at least, is robust, a blessing amidst the chaos, allowing me to savor small delights where I can. The fog of cognitive difficulties makes me feel detached from myself, a bystander in my own life. I yearn for clarity, for the sharp focus I once had. But the clouded mind brings moments of peace too, where I can escape the harsh realities, finding solace in daydreams and memories. This isn't living, I know. Each day is a struggle to find the light—the hope that tomorrow will be better. I grasp at strands of optimism, determined to believe my situation will improve. I want to feel the sun on my face, to run, to walk without pain. I want to embrace life fully again and find joy in the world. This is my new normal, a battle that demands every ounce of my being. Yet, I know that even on the darkest days, there is a chance for things to shift, for the clouds to part and hope to shine through. I cling to that hope—it's the fuel that keeps me going.

Table E8: The two generated narratives with the overall lowest scores for quality and accuracy. The narrative by `Command-r-v01` received a mean score of 2.33 out of 5 for relevance, 3.33 for coherence and empathy, 3.00 for engagement, and 0.17 for accuracy. The narrative by `Bio-Medical-Llama-3-8B` received a mean score of 3.33 for relevance, 3.67 for coherence, 2.00 for empathy, 2.33 for engagement, and 0.21 for accuracy. Quality and accuracy are weighed equally to determine the narratives with the overall lowest scores.

Model and narrative
`Bio-Medical-Llama-3-8B`. As I wake up each morning, I try to shake off the fatigue that has become all too familiar. It's not just the tiredness that comes with aging; it's the exhaustion that cancer brings. Despite this, I make it a point to get out of bed and start my day. I head to the golf course, a place where I find solace and peace. Watching the sun rise over the greens is a beautiful sight, and for a brief moment, I forget about my cancer. After my morning golf session, I head to work at the local horeca. I love interacting with the customers and making sure they have a great dining experience. My colleagues are like family to me, and we all support each other through the tough times. On good days, I can forget about my cancer and just enjoy the simple things in life. However, there are days when my cancer gets the better of me. When I'm feeling particularly fatigued, I may not be able to muster up the energy to go to work or play golf. These days are difficult, and I find myself stuck in bed, feeling sorry for myself. But even on these days, I try to stay positive and focus on the things I can do. My appetite has also changed since my diagnosis. Some days I have no interest in food, and other days I can eat a whole meal by myself. It's frustrating, but I've learned to roll with the punches. My friends and family are always encouraging me to eat, but I know my body best. I listen to it and feed it when it needs nourishment. Cancer has also taken a toll on my social life. Some days I just want to be alone and not think about anything. But other days, I force myself to reach out to friends and family. It's not always easy, but I know I need their support to get through this. Looking back on my day, I realize that cancer has changed many aspects of my life. While some days are better than others, I try to stay positive and focus on the things I can control. I know I'll have good days and bad days, but I'm determined to live my life to the fullest, even with cancer.

Table E8: (continued)