Investigating Wit, Creativity, and Detectability of Large Language Models in Domain-Specific Writing Style Adaptation of Reddit's Showerthoughts

Tolga Buz*, **Benjamin Frost***, **Nikola Genchev***,
Moritz Schneider, **Lucie-Aimée Kaffee**, **Gerard de Melo**
Hasso Plattner Institute / University of Potsdam, Germany
tolga.buz@hpi.de, gdm@demelo.org

Abstract

Recent Large Language Models (LLMs) have shown the ability to generate content that is difficult or impossible to distinguish from human writing. We investigate the ability of differently-sized LLMs to replicate human writing style in short, creative texts in the domain of *Showerthoughts*, thoughts that may occur during mundane activities. We compare GPT-2 and GPT-Neo fine-tuned on Reddit data as well as GPT-3.5 invoked in a zero-shot manner, against human-authored texts. We measure human preference on the texts across the specific dimensions that account for the quality of creative, witty texts. Additionally, we compare the ability of humans versus fine-tuned RoBERTa classifiers to detect AI-generated texts. We conclude that human evaluators rate the generated texts slightly worse on average regarding their creative quality, but they are unable to reliably distinguish between human-written and AI-generated texts. We further provide a dataset for creative, witty text generation based on Reddit Showerthoughts posts.

1 Introduction

As Large Language Models (LLMs) continue to advance, it becomes increasingly challenging for humans to distinguish AI-generated and human-written text. Generated text may appear surprisingly convincing, inciting debates whether new forms of evaluating models are necessary (Sejnowski, 2023). The high quality of LLM outputs can benefit diverse use cases, while also increasing the risk of enabling more sophisticated spam, misinformation, and hate speech bots (Manduchi et al., 2024). LLMs are known to master various aspects of grammar and basic semantics. Yet, one goal that still has proven non-trivial using LLMs is that of generating creative text (Chakrabarty et al., 2023a), especially in the realm of humour (Jentzsch and Kersting, 2023).

We seek to understand the ability of differently-sized LLMs to replicate human writing style in short and creative texts as shared in the *Showerthoughts* community on Reddit, which exhibits humour, cleverness, and creativity – often in a single sentence. The Showerthoughts community (Reddit's 11th largest) provides a unique dataset of short texts with a characteristic writing style drawing from general creative qualities. To understand how well models of different sizes can replicate such witty Reddit posts, we fine-tuned two LLMs, GPT-2 (Medium) and GPT-Neo, on posts from this online community. Additionally, we used GPT-3.5-turbo as a zero-shot model, i.e., without additional fine-tuning for our specific task. We evaluated how well the AI-generated texts emulate the style of Showerthoughts. To this end, we employed a mixed-method approach: We compare genuine, human-authored posts with generated Showerthoughts based on various lexical characteristics as well as in their similarity in sentence embeddings. Furthermore, we conducted a human evaluation study to assess the human evaluators' perception of the creative quality (specifically, logical validity, creativity, humour, and cleverness) and to measure how easily AI-generated texts can be detected.

We find that participants cannot reliably detect AI-generated texts, as the LLMs come close to human-level quality. Generating humour remains a challenging task, but shows a promising future for the generation of short, witty, and creative statements. We find that a machine learning (ML) classifier, trained on Showerthoughts, succeeds at robustly distinguishing human-authored from AI-written text. Thus, there remains potential for current AI-generated content to be identified, even in the ambiguous realm of humour and creative text.

We summarize our contributions in this paper as follows: (1) A new dataset for creative, witty text generation based on Reddit Showerthoughts

*Equal contribution

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 291 - 307

posts.[1] (2) Experiments with three different models for the generation of creative, witty text. (3) Evaluation of human perception of creative language generation through a survey. (4) Experiments on automated authorship identification of the text as human-written or AI-generated.

2 Background and Related Work

Reddit and Showerthoughts Reddit is a social media platform that is organized in communities called *subreddits*, which exist for a plethora of topics – all written, curated, voted, and commented on by the community. This provides a diverse and valuable research subject; each subreddit is characterized by a distinct writing style and type of content (Agrawal et al., 2022; Buz et al., 2024).

Our work is centered on the r/Showerthoughts subreddit[2], which defines *Showerthought* as "a loose term that applies to the types of thoughts you might have while carrying out a routine task like showering, driving, or daydreaming. At their best, Showerthoughts are universally relatable and find the amusing/interesting within the mundane." In general, popular Showerthoughts exhibit wit (or cleverness), creativity, and sometimes humour, which come from the realization of matters that lie in everyday life's banality, which are well thought out but tend to go unnoticed. They condense various intellectual qualities into short texts that often allude to a deeper context – these qualities can be facilitators of a text's success in various other settings, including posting on social media or copywriting for marketing purposes. One of the community's most successful post goes as follows: "When you're a kid, you don't realize you're also watching your mom and dad grow up."[3]

To the best of our knowledge, there is only one related paper focused on Showerthoughts, which covers a neuro-scientific perspective (Crawford, 2020). Limited research exists that uses Showerthoughts data among other subreddits, but on completely different topics, e.g., detection of suicidal thoughts (Aladağ et al., 2018), predicting conversations (Kim et al., 2023), or changes of the community (Lin et al., 2017). Our work is the first to analyse the texts that are shared in this community from a perspective of computational linguistics and the first to publish a Showerthoughts dataset.

[1]Dataset accessible via our GitHub repository.
[2]www.reddit.com/r/Showerthoughts
[3]Accessible via https://www.reddit.com/awd10u/

Creative Quality in Natural Language Generation Early work on computational creativity found that while computers can aid in the creative process, it has long remained difficult to achieve novelty and quality with such systems (Gervás, 2009). More recent LLMs possess a remarkable ability to produce entirely novel content, but Chakrabarty et al. (2022a) find that they have limited capabilities w.r.t. figurative language, and that full stories generated by LLMs seem to be of far inferior quality compared to those written by professional authors (Chakrabarty et al., 2023a). Further, popular LLMs such as ChatGPT have been found to be subpar at writing creative and humourous content such as jokes (Jentzsch and Kersting, 2023). For many creative tasks, such as writing convincing poems, human intervention may be needed to create high-quality text (Chakrabarty et al., 2022b), and the temperature hyperparameter may have a significant impact on the creativity of LLM-generated texts (Davis et al., 2024). AI-assisted writing may lead to improved results (Roemmele, 2021) and LLMs have been perceived as writing collaborators by professional writers (Chakrabarty et al., 2023b). However, it is yet to be seen how the generation of creative, witty text without human intervention can be improved to agree with human preferences.

Authorship Identification There have been significant advancements in LLMs generating grammatically correct sentences adhering to semantic rules, even purportedly attaining human levels (Köbis and Mossink, 2021; Clark et al., 2021). This presents opportunities in areas such as accessibility of information and education, and enhanced productivity (Dwivedi et al., 2023; Noy and Zhang, 2023). However, it also poses a threat to the credibility of information (Kreps et al., 2022; Kumar and Shah, 2018), especially as social media users often fail to detect bots (Kenny et al., 2022), while such bots continue to evolve and spread misinformation (Abokhodair et al., 2015; Shao et al., 2018). Indeed, Ippolito et al. (2020) found that even trained participants struggle to identify AI-generated texts. Köbis and Mossink (2021) further found that while completely random texts could be detected, cherry-picked texts could not be distinguished by humans. The model size used to generate texts affects participants' performance – both studies used smaller models (GPT-2 with 355M, 774M, and 1.5B parameters, respectively), whereas

participants confronted with never models such as GPT-3 performed significantly worse in a similar study (Clark et al., 2021; Brown et al., 2020). With larger model sizes, humans require more time to decide, and their accuracy declines (Brown et al., 2020). In very recent work, Chen and Shu (2024) find that LLM-generated misinformation can be more deceptive than when written by human authors, and Muñoz-Ortiz et al. (2023) identify measurable differences between AI-generated and human-written texts.

As LLMs advance rapidly, it becomes crucial to understand what type of generated content humans can detect and how to detect generated content automatically. For automatic authorship identification, Wani and Jabin (2017) use ML classifiers to detect bots. Ippolito et al. (2020) use a fine-tuned BERT-based binary classifier to label texts as human-written or AI-generated. However, their model lacks generalizability – when trained on top-k samples and evaluated on non-truncated random samples, the model only achieves 43.8% accuracy. The sharp increase in discussions about misuse and plagiarism using tools such as ChatGPT has shifted researchers' focus on this area, e.g., Mitchell et al. (2023) proposed DetectGPT, a zero-shot model for detecting AI-generated text, and Deng et al. (2023) proposed a Bayesian Surrogate Model, claiming to outperform DetectGPT. Tang et al. (2024) provide an overview of further detection techniques.

3 Data Compilation

To create the Showerthoughts dataset, we used the publicly available Pushshift API (Clark et al., 2021; Brown et al., 2020) to extract submissions from the Showerthoughts subreddit from April 2020 to November 2022, resulting in an initial collection of 1.3 million posts.[4] We discard posts that have been deleted or removed (often due to rule violation) as well as those that contain images or additional explanations in their body text (as the community's rules require the full Showerthought to be contained in the title). Accordingly, we only use each post's title for our experiments, resulting in a dataset of 411,189 Showerthoughts. An analysis of the most frequent choices of words reveals that they are often about people, life, common objects, and the world in general. A frequent word analysis indicates that they often compare things using, e.g., "more", "other", "old", "good".

In order to obtain a ground truth about the lexical characteristics of the dataset and later compare them with the generated texts, we conducted several tests on 5,000 randomly selected examples, focusing on sentence complexity, length, grammar, and vocabulary, the results of which are summarized in Table 1 (in the first row 'Genuine'). The complexity score is based on the Flesch-Kincaid grade level, which quantifies a text's complexity based on the number of words per sentence and syllables per word (Kincaid et al., 1975). For example, a score of 7.0 indicates that a 7th-grade student (or a person with at least seven years of education) would typically be able to read and understand the respective text.[5]

4 Experimental Setup

In the following, we detail our experimental setup for addressing our three research questions. We explain our process for generating Reddit Showerthoughts-like texts with differently sized selected LLMs. These texts are subsequently evaluated through a survey, assessing several textual aspects. Additionally, we compare the ability of humans and fine-tuned BERT-based classifiers in detecting originality. An overview of this experimental setup is given in Figure 1.

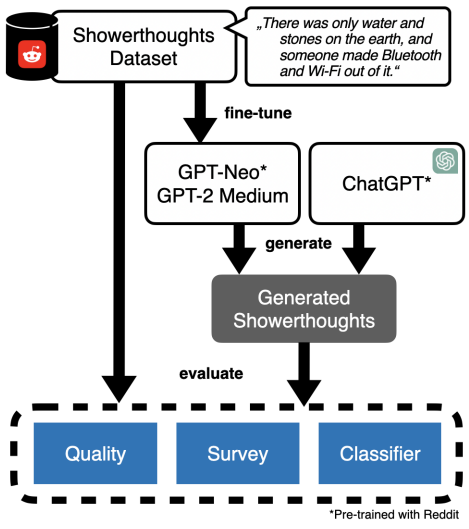


Figure 1: Overview of our experimental setup

[4] In mid 2023, Reddit changed their API guidelines, forcing Pushshift to restrict its access to Reddit moderators only. Our datasets were collected before this change occurred.

[5] Tests are conducted with the `textstat`, `language_tool_python`, and `nltk` libraries.

4.1 LLM Fine-Tuning and Prompting

We consider two setups for the generation of Showerthoughts; (1) two models of different sizes are fine-tuned; (2) ChatGPT (based on GPT-3.5-turbo) is invoked to generate Showerthoughts in a zero-shot setting.

Fine-tuning GPT-2 and GPT-Neo For the fine-tuned models, we select GPT-2 Medium (355M parameters) and GPT-Neo (2.7B parameters) and fine-tune them on the aforementioned Showerthoughts dataset. To later be able to induce the models to generate Showerthoughts, each instance is wrapped around two previously unseen tokens, `<|showerthought|>` and `<|endoftext|>`. These serve as prompt and end-of-text markers, respectively, during generation. We use the standard parameters for text generation for both models, including a temperature value of 0.9.

GPT-2 Medium[6] is a unidirectional causal language model that generates text sequences, using 355 million parameters. This was the smallest LLM still able to generate sensible results in our initial evaluation during LLM selection. We use AdamW for optimization, the `GPT2Tokenizer`, a maximum learning rate of 3×10^{-5} with 5,000 warm-up steps, a batch size of 16, and train the model for five epochs on the task of next token prediction.

GPT-Neo is an architecturally upgraded model compared to GPT-2 that closely resembles GPT-3, with 2.7 billion parameters and trained on the Pile dataset (Gao et al., 2020). We selected the same hyperparameters as for GPT-2 besides using Adafactor optimization, which provides manual control over the learning rate and has better memory efficiency (Shazeer and Stern, 2018). We used a learning rate of 2×10^{-5}, which is reduced to 7×10^{-6} over five epochs, and a batch size of 32.

Zero-shot Text Generation with ChatGPT In initial experiments, we found that a basic prompt (*"Please generate 10 Showerthoughts"*) results in repetition of content and structure in generated texts, in accordance with the findings of Jentzsch and Kersting (2023). We therefore extended the prompt by including a definition of Showerthoughts, alongside instructions for enhancing wit, creativity, and humour, and varying sentence structure. This resulted in the following prompt:

"Please generate 100 Showerthoughts, which are inspired by the Reddit community r/Showerthoughts. Vary the sentence structure between the different sentences, and try to be clever, creative, and funny. The Showerthoughts should be relatable and connected to things that people might encounter during mundane tasks."

This process was repeated 50 times to sample a total of 5,000 Showerthoughts. We use the standard settings for text generation, including a temperature value of 0.7.

4.2 Survey of Human Preferences

We evaluated the results of the text generation models by means of a survey. The participants were randomly split into two groups to evaluate a larger number of Showerthoughts while ensuring an adequate number of responses per Showerthought and a reasonable completion time (around 25 minutes). Each group evaluated 15 human-written and ten AI-generated Showerthoughts, each from GPT-2, GPT-Neo, and ChatGPT. Participants were not informed about the distribution of the sources and received the texts in a random order to prevent evaluation bias. The Showerthoughts were selected randomly and manually filtered to exclude posts harboring vulgarity or a "not safe for work" (NSFW) topic.

The survey starts with a briefing on Reddit and r/Showerthoughts, and we informed participants that they will evaluate 45 Showerthoughts, some of which are written by humans and some generated by LLMs. We further ask demographic questions, including age group and the level of experience with Reddit, Showerthoughts, and Machine Learning on a five-point scale. Then, participants were asked to evaluate a series of 45 Showerthoughts by rating along six dimensions (each on a six-point Likert scale): (1) "I like this Showerthought", (2) "It makes a true/valid/logical statement", (3) "It is creative", (4) "It is funny", (5) "It is clever", and (6) "I believe this Showerthought has been written by a real person". These criteria were selected to capture the quality of a Showerthought from diverse angles, and are also applicable to comparable short texts such as social media posts and marketing texts. For evaluation, we consider the average scores of the selected Likert scale from 1 (lowest) to 6 (highest). This method is widely used, e.g., in Tang et al. (2021). Finally, the participants could optionally provide a free-text explanation or reasoning on how they decided.

[6] Accessible via https://huggingface.co/gpt2-medium

4.3 Authorship Identification

As a counterpart to the human evaluators on the task of authorship identification, we fine-tuned a total of four RoBERTa-based models[7] (Liu et al., 2019) for binary classification of each input Showerthought as either human-written or AI-generated. For the training and testing of the three LLM-specific RoBERTa classifiers, we used 10,000 randomly selected Showerthoughts per class (i.e., genuine, generated) for GPT-2 and GPT-Neo, and 5,000 examples for the ChatGPT version (due to the smaller generated dataset size). In addition, we trained and tested another RoBERTa classifier on a combined set of 15,000 examples per class (i.e., 5,000 per LLM source). All datasets were randomly split at a 80–20 ratio for training and testing. We assessed the classifiers in three setups; (1) evaluating the three LLMs' outputs compared to human-written (genuine) text separately; (2) evaluating all three LLMs' outputs combined compared to human-written text; (3) training the classifier on one LLM's outputs (GPT-Neo) and evaluating it on another LLM's outputs (GPT-2, ChatGPT, and all combined).

All versions of the classifier were trained with the tokenizer of RoBERTa-Base, AdamW optimization, a learning rate of 2×10^{-5}, batch size of 32, and a linear scheduler with 300 warm-up steps. To compute the loss for a given prediction, the model receives the tokenized Showerthought and the corresponding label indicating whether the Showerthought was genuine or generated.

5 Results

This section presents our experimental results. Section 5.1 compares lexical characteristics, showing that the LLMs come close to human quality. Next, Section 5.2 explores the survey results, providing insights into crucial Showerthought attributes such as logical validity and creativity. Lastly, Section 5.3 reports on our authorship identification, including patterns to distinguish between human-written and AI-generated Showerthoughts.

5.1 Characteristics of Generated Showerthoughts

To assess the quality and similarity of generated to original Showerthoughts, we apply the linguistic metrics described in Section 3 to the AI-generated Showerthoughts utilizing 5,000 random samples per source (for ChatGPT we use all 5,000 texts generated). Table 1 shows that human-written (genuine) Showerthoughts have a larger vocabulary, are slightly more complex, and contain more difficult words and grammar mistakes. Based on these metrics, GPT-Neo's generated texts are closer to genuine texts compared to the significantly smaller GPT-2. ChatGPT ranks closest to the human reference regarding average complexity and length, slightly behind GPT-Neo regarding vocabulary size, but farthest away from the reference in terms of difficult words and grammar mistakes. We find that the models produce a negligible amount of duplicate Showerthoughts (GPT-2: 13 of 10,000, GPT-Neo: 162 of 10,000, ChatGPT: 6 of 5,000).

Source	Genuine	GPT-Neo	GPT-2	ChatGPT
Compl.[1]	7.4 ± 3.4	6.8 ± 3.0	6.3 ± 2.7	6.9 ± 2.4
Length[1]	81 ± 38	88 ± 39	87 ± 33	81 ± 21
Vocab.[2]	13,000	8,700	4,900	7,200
Diffic.[3]	1.2	0.7	0.4	0.36
Errors[3]	0.3	0.2	0.1	0.05

[1] Mean linguistic complexity (Flesch-Kincaid grade level) and length with standard deviation.
[2] Vocabulary size in number of unique words.
[3] Number of difficult words and grammatical errors per sentence.

Table 1: Comparison of common lexical characteristics (based on 5,000 random samples per source)

Comparison of Sentence Embeddings How semantically diverse are Showerthoughts and are our LLMs able to match this diversity? To answer this, we employ sentence embeddings[8] for comparing the similarity between human-written and AI-generated content, and to measure the linguistic distance to texts from other subreddits. We have reviewed the embeddings of 1,000 randomly sampled Showerthoughts per source visualized with the t-SNE algorithm (Van der Maaten and Hinton, 2008); GPT-2 and GPT-Neo produce more diverse texts than zero-shot ChatGPT, which matches human-written Showerthoughts based on their output distributing across the same semantic clusters as the human-written texts (Figure 2, in Appendix). When comparing these embeddings to 1,000 randomly selected titles from different, similarly large and popular subreddits, we find that every subreddit has a distinct focus, and the generated and genuine Showerthoughts being in the same cluster indicates

[7] Specifically: *RoBERTaForSequenceClassification*.

[8] SBERT embeddings in their default, pre-trained configuration (`all-MiniLM-L6-v2`)

that the models are successful in replicating the distinct writing of each subreddit (Figure 3, in Appendix).

5.2 Survey Results

A total of 56 human evaluators took our survey (25 participants in Group A and 31 in Group B), resulting in an accumulated 2,520 ratings for the full set of 90 Showerthoughts and an average of 28 ratings per item, as each group reviewed a completely different set of 45 texts.

Demographics of Survey Participants The participants' demographics are influenced by the channels the survey was shared in: The majority of the participants are younger than 30 years old, with 8.5% above 30 years. 89.4% of respondents have some degree of machine learning (ML) experience, 42.6% have trained an ML model at least once, and some of these even work with ML models daily. Only 10.6% indicated little to no experience with ML. 53.1% of participants rarely or never visit Reddit, while the rest visit monthly (8.5%), weekly (38.3%), or daily (27.7%). 31.2% had never heard of r/Showerthoughts before, while 68.7% visited the community at least once in the past – 16.6% are subscribed and follow it regularly, with 6.2% even occasionally engaging in the community.

It is clear that this demographic distribution is not representative for the broader population, but a result of the distribution channels used for the survey: the professional and university networks of the authors. From a statistical perspective, this is likely to introduce a bias – however, we find it highly interesting to study this group of individuals nonetheless, as many are experienced with ML and approximately half are familiar with Reddit, which we hypothesize to potentially improve their abilities.

Source	Genuine	GPT-2	GPT-Neo	ChatGPT
Score	**3.71**	2.42	3.40	3.23
Log. Val.	**4.20**	3.10	3.96	3.55
Creativity	**3.63**	2.42	3.23	3.45
Humour	**3.18**	2.10	2.74	2.85
Cleverness	**3.41**	2.19	3.15	3.07

Table 2: Mean score (on a six-point scale) for the Showerthought quality criteria (Log. Val. = Logical Validity); best score bold, best model underlined

Overview of Showerthought Ratings Table 2 displays the average response scores for the first five evaluation criteria. None of the LLMs is able to beat or match the scores of human-written Showerthoughts, but some of them get remarkably close.

Among the models, GPT-Neo achieves the best ratings for general score, logical validity, and cleverness, while ChatGPT (based on GPT-3.5-turbo) performs better on creativity and humour. It appears that the general ability to write a convincing, logical, and clever Showerthought can be learned in fine-tuning, but more abstract abilities like creativity and humour improve with model size.

The smallest model, GPT-2, performs the worst, consistently short of human-written Showerthoughts, exhibiting an approximately 30% worse performance. GPT-Neo and ChatGPT achieve a much smaller margin with an overall average disparity of 6% and 7%, respectively. The evaluators consistently prefer human-written texts – however, the margins are small and this does not necessarily have implications for the task of authorship identification, as we show below.

Manual Authorship Identification From the survey responses regarding authorship of a text, we consider answers between 1 and 3 as a vote for AI-generated, and answers between 4 and 6 as a vote for human-written text. Table 3 displays the average accuracy of the survey's participants in correctly identifying each Showerthought's source. For a more granular evaluation, we additionally display the responses by the participants' experience in Reddit, machine learning (ML), and Showerthoughts.[9]

We find that the survey participants were not able to consistently identify whether a Showerthought was human-written or AI-generated; Between all human-written (genuine) and GPT-2, GPT-Neo, and ChatGPT generated Showerthoughts the survey participants were only able to correctly identify 63.8%, 73.1%, 48.1%, and 46.2%, respectively. For GPT-Neo and ChatGPT, this is worse than (balanced) random guessing, i.e., a strategy that would choose one of the two classes in 50% of cases. This indicates that GPT-Neo and ChatGPT already generate Showerthoughts sufficiently convincing to mislead human evaluators. Experience with Reddit and Showerthoughts improves the participants' ability to identify human-written Show-

[9]The participants were considered 'experienced' in one of the given categories if they chose one of the top two answers (e.g., visiting Reddit 'Weekly' or 'Daily') and 'unexperienced' if they chose one of the bottom two answers (e.g., visiting Reddit 'Never' or 'Rarely').

	Overall	Reddit Experience		ML Experience		Showerthoughts Experience	
Model		Yes	No	Yes	No	Yes	No
Genuine	63.8 %	71.3 %	60.2 %	63.2 %	62.3 %	81.6 %	62.0 %
GPT-2	73.1 %	71.3 %	72.2 %	74.0 %	74.0 %	60.0 %	72.4 %
GPT-Neo	48.1 %	49.0 %	46.6 %	48.0 %	53.5 %	55.0 %	45.7 %
ChatGPT	46.2 %	43.9 %	45.1 %	46.8 %	44.0 %	42.5 %	44.3 %
No. Participants	56	21	30	25	7	5	45

Table 3: Survey participants' accuracy in correctly identifying the Showerthought's source

		Prec.	Rec.	F1	Support
GPT-2	Generated	0.91	1.00	0.95	2,000
	Genuine	1.00	0.90	0.95	2,000
	Accuracy			0.95	4,000
	Average	0.96	0.95	0.95	4,000
GPT-Neo	Generated	0.84	0.99	0.91	2,000
	Genuine	0.99	0.82	0.90	2,000
	Accuracy			0.90	4,000
	Average	0.92	0.90	0.90	4,000
ChatGPT	Generated	0.91	0.99	0.95	400
	Genuine	0.99	0.91	0.94	400
	Accuracy			0.95	800
	Average	0.95	0.95	0.95	800
Combined	Generated	0.82	0.95	0.88	3,000
	Genuine	0.94	0.79	0.86	3,000
	Accuracy			0.87	6,000
	Average	0.88	0.87	0.87	6,000

Table 4: Precision, Recall, F1, and Support of the RoBERTa models trained for Showerthoughts authorship identification (LLM-specific models and one combined model for all LLMs)

erthoughts, but does not improve their ability to detect AI-generated texts consistently.

To investigate whether evaluators are more accurate with higher confidence, we evaluated high-confidence answers only (i.e., 1 – 2 and 4 – 6). However, detection accuracy did not improve. In these cases GPT-2 was detected with an accuracy of 79.6%, while there were only small improvements in detecting the other sources. The detection accuracy regarding GPT-Neo and ChatGPT remained below the random-guess baseline. Similar to the overall results, experience with Reddit or Showerthoughts only helped in identifying genuine texts. This shows that independent of their size GPT-Neo and ChatGPT are able to mislead evaluators with the quality of their generated texts.

Participants' Reasoning for Detecting AI-Generated Texts At the end of the survey, participants could add explanations for their evaluation. Within the 42 responses, the primary factors were: illogical statements, common sense, good grammar, lack of humour / depth / creativity, and repetitive word or syntax usage. Endowing machines with commonsense knowledge has been a long-standing goal in AI (Tandon et al., 2017), which LLMs address to a significant degree. The finding that 'good grammar' was frequently mentioned is noteworthy, as many participants believed that machines excel at grammar while errors indicate human authorship. These findings are consistent with prior research by Dugan et al. (2022), who identified similar factors as the most commonly cited indicators of AI-generated content.

5.3 Automated Authorship Identification

This section presents the evaluation results of the four different RoBERTa classifiers introduced in Section 4.3 – three LLM-specific classifiers and one trained on the combined texts of all three models. The classification reports presented in Table 4 show that the classifiers trained per model achieve an overall accuracy ranging from 90% to 95%, with the single model trained for all LLMs scoring an accuracy of 87% (Table 4). Across all classifiers, recall for LLM-generated instances approaches 100% with lower precision, while precision for the genuine human-authored class is nearly perfect but with lower recall. These findings indicate the following: (1) These classifiers outperform human evaluators on authorship identification.[10], (2) The

[10]Note: While human evaluators receive a more general instruction at the beginning of the survey, the classification models are fine-tuned for the task. Nonetheless, we consider this a realistic setup, as almost 70% of the evaluators have responded to have prior experience with the Showerthoughts community. For future work, human evaluators could be presented with human-written and AI-generated examples at the beginning of the survey.

classifiers consistently misclassify a portion of genuine Showerthoughts as generated, which are either lower-quality examples or similar to generated texts in some regard. (3) The models perform well in detecting the AI-generated texts, with the combined RoBERTa model achieving an average F1 score of 0.87. (4) Current (GPT-based) language models, independent of their size, appear to utilize similarly transparent techniques for language generation and are therefore similarly easy to detect for an ML classifier, even when trained on a different GPT-based model.

In an additional experiment, we trained a classifier to distinguish texts of GPT-Neo from genuine ones but evaluate its performance on texts of the other LLMs. The results in Table 5 show that the classifier's average performance on the texts of other models can achieve a relatively high value of 0.86 when a single model's texts are utilized for evaluation. However, the results are significantly worse when texts of various models, of which most were not part of the training, are included for evaluation, suggesting fine-tuning with texts from multiple LLMs for better detection performance.

Our evaluation of the fine-tuned RoBERTa models shows that none of the classifiers attain 100% accuracy, emphasizing caution when using detection tools, particularly in cases with serious consequences such as academic failure or job loss. In a real-world setting, the specific LLM invoked to generate and spread texts will likely be unknown, and, therefore, cannot provide training samples, which requires robust generalizable classifiers and non-GPT-based LLMs – important questions requiring investigation in future work. Nonetheless, our results suggest that the models have learned patterns that strongly indicate whether a given Showerthought is AI-generated, which proves valuable for evaluating the tokens and patterns that contribute the most to the classification results, which we do in the following section.

Tokens with Greatest Contribution towards Class Prediction We use the LLM explainability library `transformers-interpret` to identify the most influential tokens per RoBERTa model. For evaluating correctly and falsely classified texts, we select the top four contributing tokens to each Showerthought's predicted class, then aggregate and normalize each token's significance relative to the dataset.

The results for the three LLMs are similar – significant contributors are (1) tokens at the beginning of a sentence, as they start with a capitalized first letter ('If', 'The', 'You' and 'We' seem to be frequent in generated texts) and (2) punctuation ('.' and ',' specifically). Punctuation and specific stop words (e.g., 'you', 'the') seem to be tokens with high attribution scores for the genuine class, indicating that a critical difference between the two classes is the placement of these tokens. ChatGPT shows slightly different top contributors, especially 'Why' and '?' – this model seems to generate questions more frequently and seems to have a unique usage of the word 'is'. Differences between ChatGPT and the other models may result from ChatGPT's pre-training data including a different subset of Reddit data and the model's much larger size.

Furthermore, our results indicate that those human-written Showerthoughts falsely classified as AI-generated by GPT-2 and GPT-Neo share the characteristics identified of generated texts, e.g., starting sentences with 'You', 'The', and 'We'. ChatGPT shows fewer distinct patterns in contributor variety and overlap between correct and incorrect human classifications. Showerthoughts mistaken as human-written ones use punctuation and blank spaces in a similar way as the genuine texts, while the misclassified human-written texts use words that may occur rarely, or seem to originate from another language. We provide more detailed results in the Appendix. In summary, RoBERTa classifiers have difficulties in cases where the characteristic writing styles of the classes overlap (especially for GPT-2 and GPT-Neo) or the misclassi-

		Prec.	Rec.	F1	Support
GPT-2	Generated	0.83	0.90	0.86	2,000
	Genuine	0.89	0.82	0.85	2,000
	Accuracy			0.86	4,000
	Average	0.86	0.86	0.86	4,000
ChatGPT	Generated	0.99	0.74	0.85	400
	Genuine	0.79	0.99	0.88	400
	Accuracy			0.86	800
	Average	0.89	0.86	0.86	800
All models	Generated	0.75	0.54	0.62	3,000
	Genuine	0.64	0.82	0.72	3,000
	Accuracy			0.68	6,000
	Average	0.69	0.68	0.67	6,000

Table 5: Evaluation of the RoBERTa model trained on on GPT-Neo's generated texts when evaluated on texts from other sources

fied Showerthought contains rarely-used or foreign words.

6 Conclusion

In this study, we demonstrate that relatively small, GPT-based LLMs can be fine-tuned to replicate the writing style of short texts of high creative quality, using the Showerthoughts subreddit as an example. While it remains to be investigated to what extent the creativity stems from observations encountered in the pretraining corpus as opposed to novel creations, we find that large numbers of diverse texts can be produced with great ease. Human raters confirm that the generated texts exhibit wit, creativity, and humour. This paves the way for diverse applications in productivity, creative work, and entertainment, and is relevant for practitioners deploying small LLMs to be cost-efficient.

We find that human evaluators rate the generated texts on average slightly lower regarding creativity, humour, cleverness. This does not seem to aid in authorship detection ("I believe this Showerthought has been written by a real person"), as we find that evaluators could not reliably distinguish AI-generated texts from human-written ones. Additionally, the quality of human-written Showerthoughts varies, with bad ones often being mislabeled as AI-generated.

Nonetheless, the possibility to abuse these models to produce spam, misinformation, or other harmful content is a growing concern. Our RoBERTa-based authorship identification classifiers performs well after fine-tuning, revealing interesting hidden patterns that help in detecting the texts generated by specific LLMs. While ML classifiers can currently detect AI-generated texts (when fine-tuned for the task), we can assume that the text generation quality of LLMs will further improve, making this task more difficult. Additionally, differently designed models may pursue other strategies for generating texts, necessitating their inclusion when training general-purpose classifiers.

Our work extends existing work that LLMs can learn to generate specific types of texts (when fine-tuned on high-quality data) to the domain of creative and witty texts, as exhibited by Showerthoughts, but not limited to those. For example, practitioners who would like to utilize such a LLM for marketing or copy-writing, could not only prompt it for general Showerthoughts about a random topic, but also add the start of a text or topic to their prompt for the LLM to complete. Alternatively, generated texts can be clustered by topic to identify the right topics for a specific use case. Simultaneously, we strongly recommend further research on detection mechanisms – while training detection models using generated texts of known LLMs and those fine-tuned on known datasets seems feasible, the task becomes more difficult when there is an exceedingly high number of LLMs to consider and even more so if the author-LLM's architecture or the training dataset is not known.

Ethics Statement

As the dataset proposed in this paper (see Section 3) is based on real user-submitted data from the Reddit Showerthoughts community, it is important to handle it with care. It should not be used to identify individuals and might contain offensive text or wrong information. This should be considered in future use of the dataset. For the survey (see Section 4.2), we manually removed inappropriate content to make it appropriate for the context of where the survey was distributed, e.g., university mailing lists. The type of survey conducted here is exempt from an ethics board review at our institution, as we have carefully designed it to be transparently described and to avoid collection of personal data.

References

Norah Abokhodair, Daisy Yoo, and David W. McDonald. 2015. Dissecting a Social Botnet: Growth, Content and Influence in Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 839–851, New York, NY, USA. Association for Computing Machinery.

Pratik Agrawal, Tolga Buz, and Gerard de Melo. 2022. WallStreetBets beyond GameStop, YOLOs, and the Moon: The unique traits of Reddit's finance communities. In *Proceedings of AMCIS 2022*. Association for Information Systems.

Ahmet Emre Aladağ, Serra Muderrisoglu, Naz Berfu Akbas, Oguzhan Zahmacioglu, and Haluk O Bingol. 2018. Detecting Suicidal Ideation on Forums: Proof-of-Concept Study. *Journal of Medical Internet Research*, 20(6):e9840.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tolga Buz, Moritz Schneider, Lucie-Aimée Kaffee, and Gerard de Melo. 2024. Highly Regarded Investors? Mining Predictive Value from the Collective Intelligence of Reddit's WallStreetBets. In *ACM Web Science Conference (Websci '24), May 21–24, 2024, Stuttgart, Germany*, page 11. ACM, New York, NY, USA.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not Rocket Science: Interpreting Figurative Language in Narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023a. Art or Artifice? Large Language Models and the False Promise of Creativity. *CoRR*, abs/2309.14556.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2023b. Creativity Support in the Age of Large Language Models: An Empirical Study Involving Emerging Writers. *arXiv preprint arXiv:2309.12570*.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022b. Help me write a poem: Instruction Tuning as a Vehicle for Collaborative Poetry Writing. *arXiv preprint arXiv:2210.13669*.

Canyu Chen and Kai Shu. 2024. Can LLM-Generated Misinformation Be Detected? *arXiv preprint arXiv:2309.13788*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text". In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Kevin Crawford. 2020. Daydreaming of Genius: Insight and the Wandering Mind. *Scientific Kenyon: The Neuroscience Edition*, 4(1):55–62.

Joshua Davis, Liesbet Van Bulck, Brigitte N Durieux, and Charlotta Lindvall. 2024. "The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research". *JMIR Hum Factors*, 11.

Zhijie Deng, Hongcheng Gao, Yibo Miao, and Hao Zhang. 2023. Efficient Detection of LLM-generated Texts with a Bayesian Surrogate Model. *arXiv preprint arXiv:2305.16617*.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2022. Real or Fake Text?: Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text. *arXiv preprint arXiv:2212.12672*.

Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71:102642.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*.

Pablo Gervás. 2009. Computational Approaches to Storytelling and Creativity. *AI Magazine*, 30(3):49–49.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Sophie F. Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. pages 325–340.

Ryan Kenny, Baruch Fischhoff, Alex Davis, Kathleen M Carley, and Casey Canfield. 2022. Duped by Bots: Why Some are Better than Others at Detecting Fake Social Media Personas. *Human factors*, page 00187208211072642.

Jinhyeon Kim, Jinyoung Han, and Daejin Choi. 2023. Predicting continuity of online conversations on Reddit. *Telematics and Informatics*, 79:101965.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Sarah Kreps, R. Miles McCain, and Miles Brundage. 2022. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science*, 9(1):104–117.

Srijan Kumar and Neil Shah. 2018. False Information on Web and Social Media: A Survey. *arXiv preprint arXiv:1804.08559*.

Nils Köbis and Luca D. Mossink. 2021. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114:106553.

Zhiyuan Lin, Niloufar Salehi, Bowen Yao, Yiqi Chen, and Michael Bernstein. 2017. Better When It Was Smaller? Community Content and Behavior After Massive Growth. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 132–141.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Laura Manduchi, Kushagra Pandey, Robert Bamler, Ryan Cotterell, Sina Däubener, Sophie Fellenz, Asja Fischer, Thomas Gärtner, Matthias Kirchler, Marius Kloft, Yingzhen Li, Christoph Lippert, Gerard de Melo, Eric Nalisnick, Björn Ommer, Rajesh Ranganath, Maja Rudolph, Karen Ullrich, Guy Van den Broeck, Julia E Vogt, Yixin Wang, Florian Wenzel, Frank Wood, Stephan Mandt, and Vincent Fortuin. 2024. On the challenges and opportunities in generative ai.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *arXiv preprint arXiv:2301.11305*.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting Linguistic Patterns in Human and LLM-Generated Text.

Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Melissa Roemmele. 2021. Inspiration through Observation: Demonstrating the Influence of Automatically Generated Text on Creative Writing.

Terrence J Sejnowski. 2023. Large Language Models and the Reverse Turing Test. *Neural computation*, 35(3):309–342.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Niket Tandon, Aparna Varde, and Gerard de Melo. 2017. Commonsense knowledge in machine intelligence. *SIGMOD Record*, 46(4):49–52.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The Science of Detecting LLM-Generated Text. *Communications of the ACM*, 67(4):50–59.

Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Thomas Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. Investigating Crowdsourcing Protocols for Evaluating the Factual Consistency of Summaries. *arXiv preprint arXiv:2109.09195*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Mudasir Ahmad Wani and Suraiya Jabin. 2017. A sneak into the Devil's Colony-Fake Profiles in Online Social Networks. *arXiv preprint arXiv:1705.09929*.

A Appendix

A.1 Simulation of Human Preference with GPT-4

We conducted an additional experiment using OpenAI's GPT-4 API investigating its ability to learn from the survey results to simulate the preferences of the human evaluators on a larger set of Showerthoughts. For this purpose, we defined a system prompt that includes a set of survey-evaluated examples and their average scores for all six categories to provide guidance for the model. To measure whether the few-shot prompting has a genuine effect and how the number of few-shot examples affects the results, we experimented with different amounts of examples, starting with three (3% of all survey items), 45 (50%), and 72 (80%), while using the rest of the survey items for testing.

We measured the coherence of GPT-4's test outputs with the Pearson Correlation metric, which shows a significant increase in correlation when increasing the number of examples shown to GPT-4 in the system prompt: after three examples (train), GPT-4's ratings obtain a Pearson correlation of 0.28 with the remaining human evaluations (test), whereas the correlation is 0.49 after 45 examples (50%), and 0.70 after 72 examples (which is a 80–20 train–test split). In order to further validate these results, we perform tenfold cross-validation using all 90 evaluated Showerthoughts, i.e., by splitting up the evaluated examples into groups of nine and using each group as a test set in a separate iteration, while all other groups are shown to the model as few-shot examples.

The system prompt is defined as follows:

> *Act like a frequent visitor of Reddit, and its r/Showerthoughts community in par-*

ticular. You participate in a scientific survey and utilize your experience to rate Showerthoughts across five dimensions: general score, validity, creativity, funniness, cleverness - with scores from 1 (low) to 6 (high). Additionally, you make a guess on a range from 1 to 6 whether the Showerthought was written by a human author (6) or generated by a language model (1). In order to learn how to score the Showerthoughts, you will receive examples, which have been rated by a team of human annotators. Your task is to rate Showerthoughts as similar to the human annotators as possible.

Here are the examples:
1 Most drivers of the Honda Fit are in fact not fit 3,8 3,4 4 4,28 3,2 4,28
...

For the evaluation of a larger set of 2,000 Showerthoughts per source, we provide all human-labelled items as examples within the system prompt to maximize the model's ability to simulate human preference.

A.2 Visualization of Sentence Embeddings

Figures 2 and 3 show the distribution of the SBERT embeddings for the different Showerthoughts compared to each other and compared to different subreddits selected for their similarity, respectively. These indicate that the fine-tuned LLMs in fact reproduce all topics that the original Showerthoughts cover, while ChatGPT is limited to a subset of the topics.

A.3 Further Survey Details

This section provides additional information on the conducted survey.

A.3.1 Participant Briefing

All survey participants were briefed with the following text:

"We are a group of students from the Hasso Plattner Institute in Potsdam who are taking part in the research seminar 'Recent Trends in AI and Deep Learning'. As part of the project, we have trained a Machine Learning model that is able to generate short texts in the style of the Reddit community 'Showerthoughts'. This survey aims to evaluate the quality of the generated texts compared to the original examples."

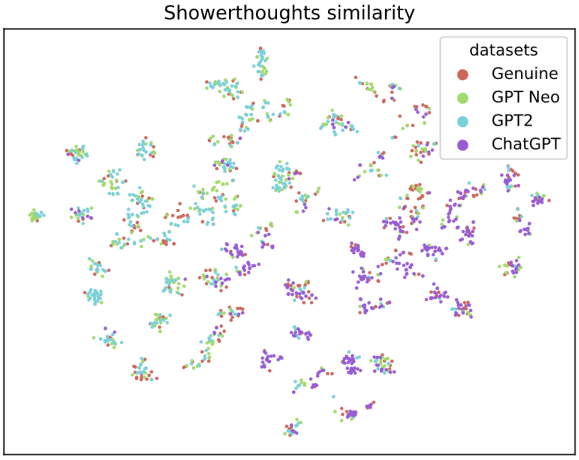


Figure 2: Semantic diversity of the different Showerthoughts datasets (t-SNE visualization of SBERT embeddings)

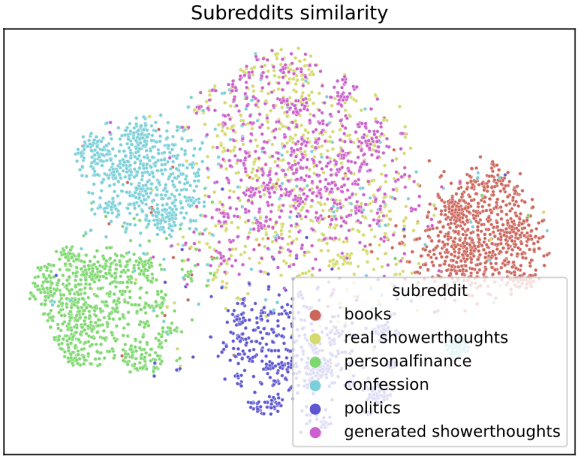


Figure 3: Comparison of genuine and generated Showerthoughts embeddings to other relevant subreddits

Definition

- **Reddit** is a social media platform that is organized in sub-communities called "subreddits". Any user can create a subreddit that revolves around any specific topic, e.g., world news, formula 1, a specific computer game, or the newest Apple iPhone. Users interested in a community can subscribe and interact within the community by posting content (self-written texts, images, videos, or links to external websites), commenting on posts, or up-/downvoting other posts and comments. Each subreddit usually has a self-defined set of rules and guidelines and is managed by a group of moderators.
- The community of **r/showerthoughts** describes itself as a "subreddit for sharing those miniature epiphanies you have that highlight the oddities within the familiar." They define a "Showerthought" as "a loose term that applies to the type of thoughts you might have while carrying out a routine task like showering, driving, or daydreaming. At their best, showerthoughts are universally relatable and find the amusing/interesting within the mundane."

Survey Setup After a few demographic questions, you will be presented with Showerthoughts, of which some are real examples from the community, and some are generated by one of three Machine Learning models (GPT-2, GPT Neo, and ChatGPT). The survey results will be anonymised and utilised only in this research project and the resulting paper. This survey consists of 5 demographics-related questions, followed by the 45 Showerthoughts, which have to be rated regarding a set of criteria each. Finally, you can optionally describe what your thinking process was like / what criteria you used to distinguish genuine from generated Showerthoughts. We estimate that the survey will take you between 20 and 30 minutes to complete.

A.3.2 Survey Questions

After the demographic questions shown in Table 6, the participants were presented a list of 45 Showerthoughts, each with six questions to answer on a six-step Likert scale (from 1= Strongly disagree to 6 = Strongly agree):

1. I like this Showerthought.
2. It makes a true/valid/logical statement.
3. It is creative.
4. It is funny.
5. It is clever.
6. I believe this Showerthought has been written by a real person.

At the end of the survey, we asked the participants a final optional question that could be answered with a free text: "When you tried to distinguish genuine from generated Showerthoughts, was there anything specific (e.g., bad grammar, or logical errors) that unveiled the generated ones?"

A.3.3 Retrieving Random Genuine Showerthoughts

In order to retrieve random genuine Showerthoughts we used an endpoint Reddit provides with its API[11]. To retrieve Showerthoughts specifically we used `GET /r/Showerthoughts/random`.

A.3.4 Statistics of Demographic Question Results

Figures 4 – 7 illustrate the survey participants' demographics and levels of familiarity with machine learning, Reddit, and the Showerthoughts subreddit.

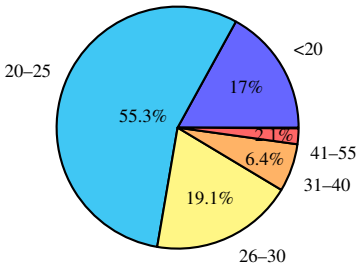


Figure 4: Age

A.3.5 Statistics on Showerthought Ratings

The box plots in Figures 8 – 13 illustrate the evaluations provided by the survey participants regarding the Showerthoughts, grouped by statement and model.

A.4 RoBERTa Interpretability Results

Figures 14 – 19 depict which tokens had the highest influence towards the predicted class.

[11] www.reddit.com/dev/api/#GET_random

Question	Answer Options
How old are you?	<20 / 20–25 / 26–30 / 31–40/ 41–55/ >55
How often do you visit Reddit?	Never / Rarely / Monthly / Weekly / Daily
Are you familiar with the r/Showerthoughts community?	No never heard of it / Visited sometime in the past / Subscribed and regularly following / Interact (post, up/downvote, or comment) rarely / Interact (post, up/downvote, or comment) regularly
How experienced are you in using Machine Learning models?	No experience / Using a product with AI or Machine Learning-based features / Played around with AI tools (e.g., ChatGPT) / Trained a ML model at least once / Working with ML models regularly

Table 6: Demographic Survey Questions and Answer Options

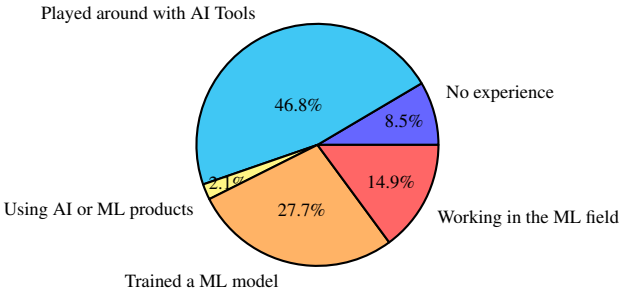


Figure 5: Experience in ML

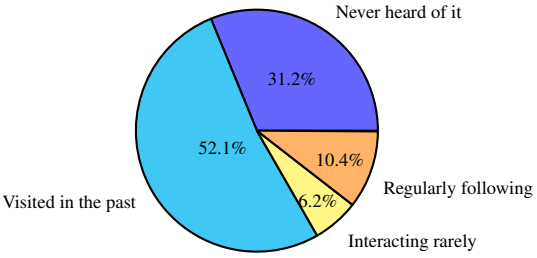


Figure 7: Familarity with Showerthoughts

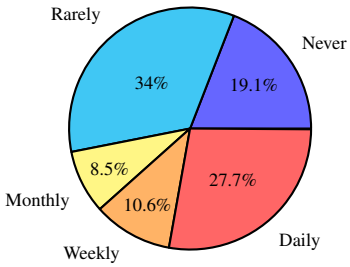


Figure 6: Reddit usage

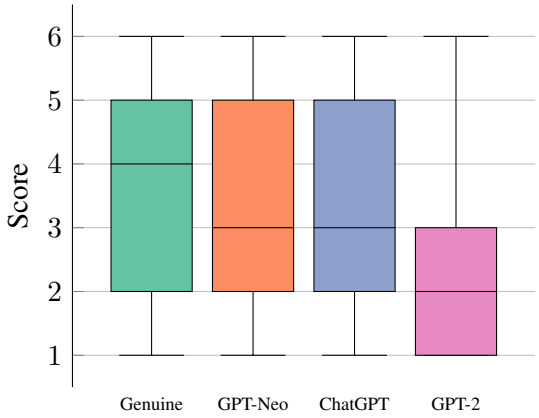


Figure 8: General Score

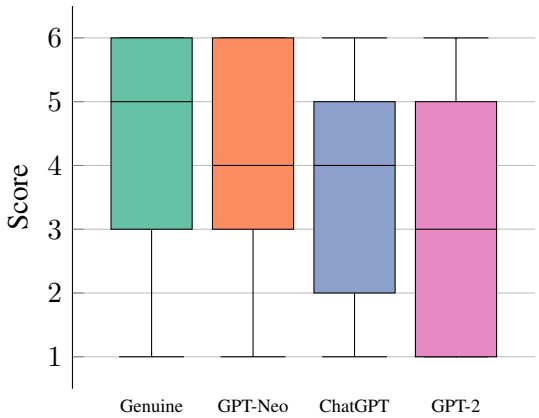


Figure 9: Logical Validity

For instance, Figures 14a, 15a, and 16a depict the most significant contributors for predicting the generated class in the training data for each of the model-specific RoBERTa classifiers.

For an additional perspective, Figures 17, 18, and 19 show the most relevant contributors for misclassified Showerthoughts, i.e., the features that influenced the respective RoBERTa classifier to predict the wrong class.

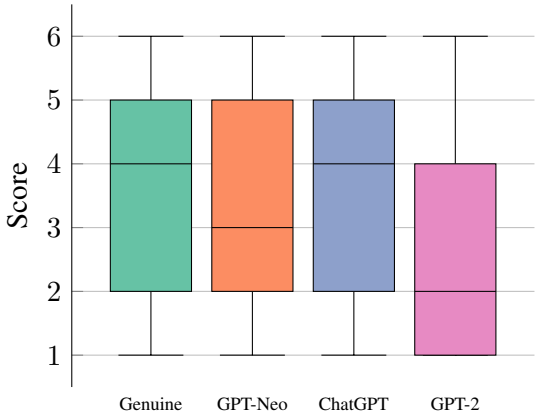


Figure 10: Creativity

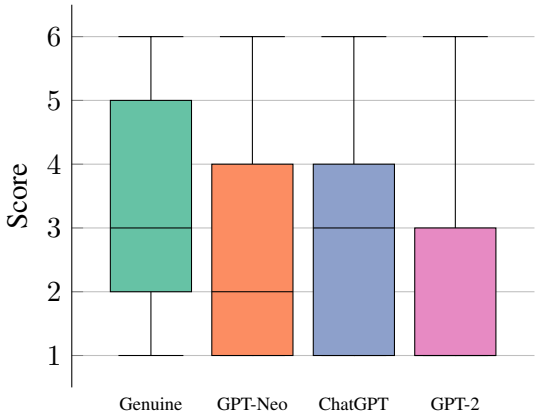


Figure 11: Funniness

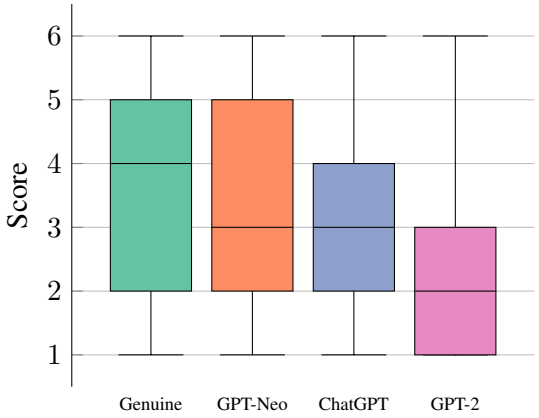


Figure 12: Cleverness

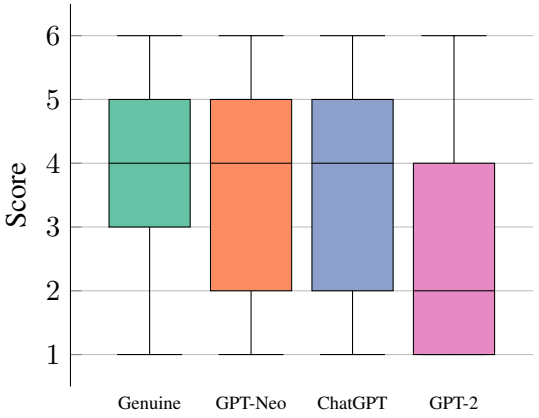
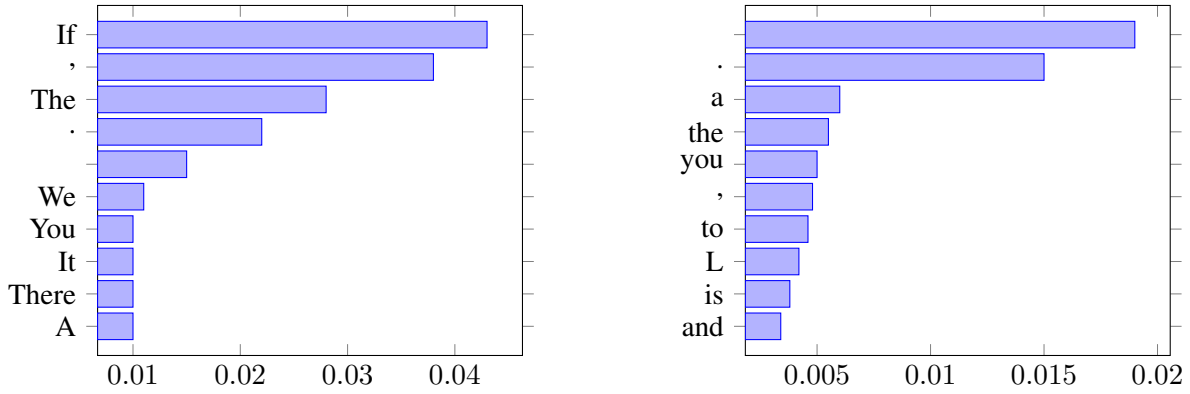


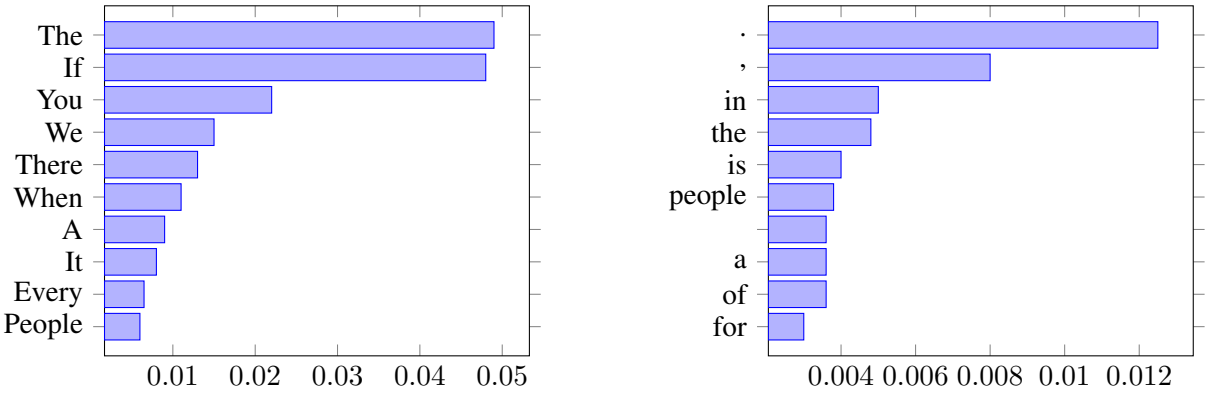
Figure 13: “I believe this Showerthought has been written by a real person”



(a) Predicted "generated" class

(b) Predicted "genuine" class

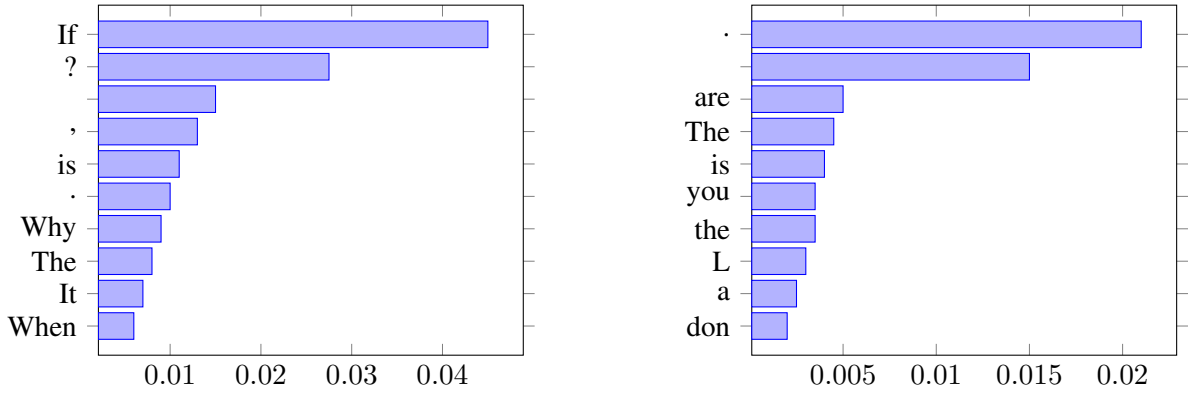
Figure 14: Tokens with highest attribution scores towards the predicted class (GPT-2)



(a) Predicted "generated" class

(b) Predicted "genuine" class

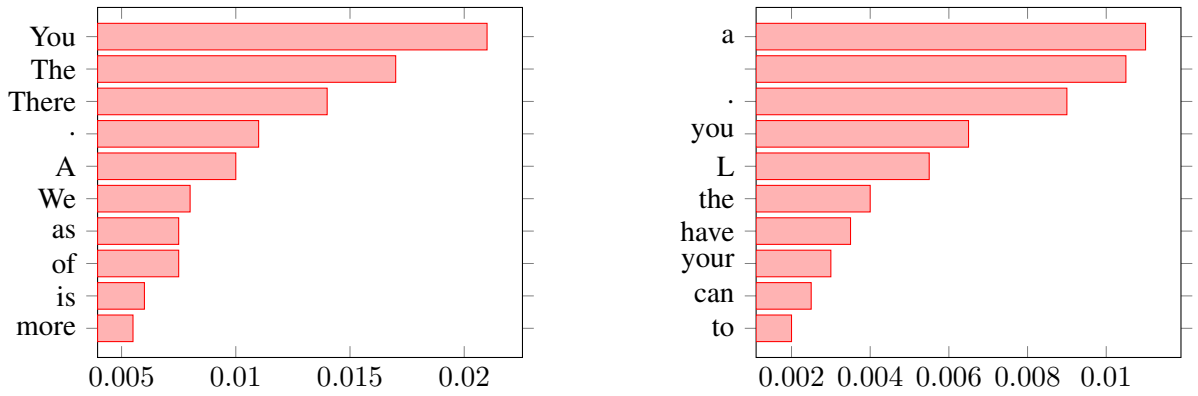
Figure 15: Tokens with highest attribution scores towards the predicted class (GPT-Neo)



(a) Predicted "generated" class

(b) Predicted "genuine" class

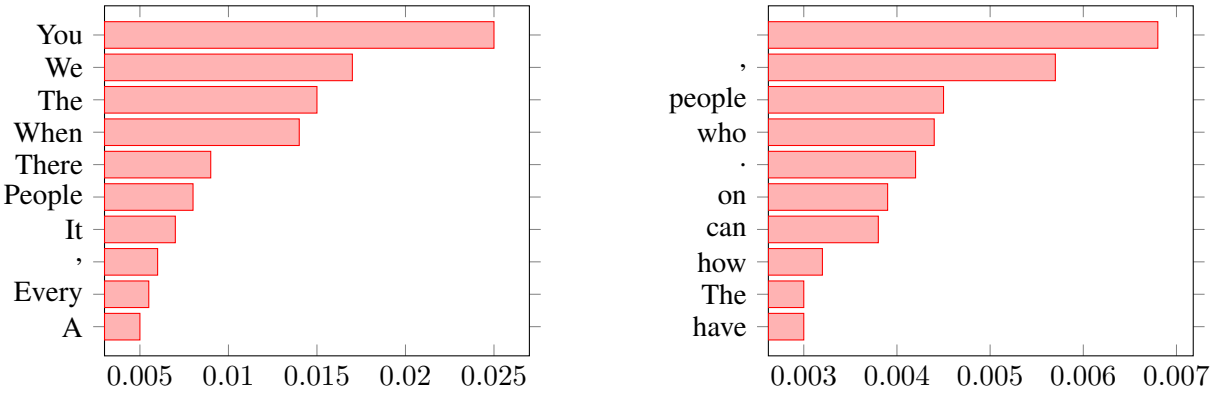
Figure 16: Tokens with highest attribution scores towards the predicted class (ChatGPT)



(a) Predicted "generated" class

(b) Predicted "genuine" class

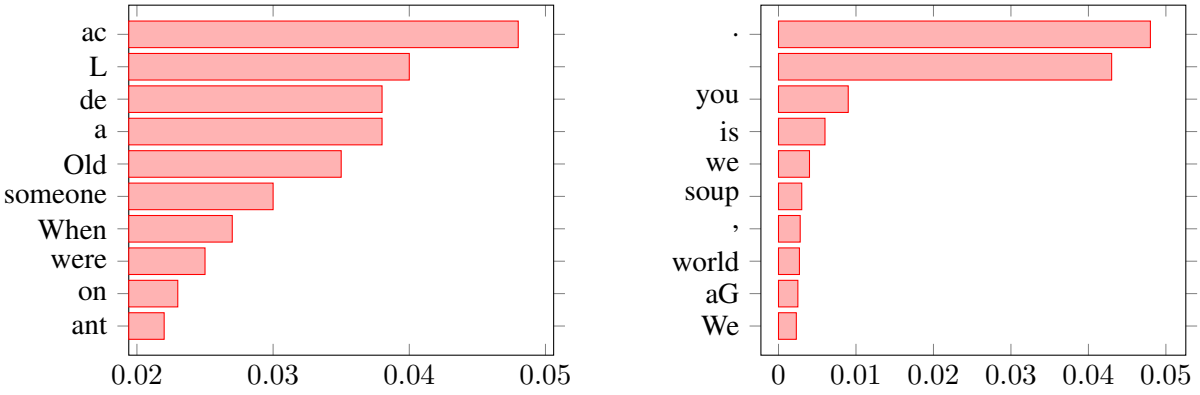
Figure 17: Tokens with highest attribution scores towards the predicted class when misclassified (GPT-2)



(a) Predicted "generated" class

(b) Predicted "genuine" class

Figure 18: Tokens with highest attribution scores towards the predicted class when misclassified (GPT-Neo)



(a) Predicted "generated" class

(b) Predicted "genuine" class

Figure 19: Tokens with highest attribution scores towards the predicted class when misclassified (ChatGPT)