Truth in the Noise: Unveiling Authentic Dementia Self-Disclosure Statements in Social Media with LLMs

Daniel Cabrera Lozoya[1], Jude P Mikal[2], Yun Leng Wong[2], Laura S Hemmy[2], and Mike Conway[1]
[1]The University of Melbourne, Australia
[2]University of Minnesota Twin Cities, USA
dcabreralozo@student.unimelb.edu.au
{jpmikal, wong0620, hemmy001}@umn.edu
mike.conway@unimelb.edu.au

Abstract

Identifying self-disclosed health diagnoses in social media data using regular expressions (e.g. "I've been diagnosed with <Disease X>") is a well-established approach for creating ad hoc cohorts of individuals with specific health conditions. However there is evidence to suggest that this method of identifying individuals is unreliable when creating cohorts for some mental health and neurodegenerative conditions. In the case of dementia, the focus of this paper, diagnostic disclosures are frequently whimsical or sardonic, rather than indicative of an authentic diagnosis or underlying disease state (e.g. "I forgot my keys again. I've got dementia!"). With this work and utilising an annotated corpus of 14,025 dementia diagnostic self-disclosure posts derived from Twitter, we leveraged LLMs to distinguish between "authentic" dementia self-disclosures and "inauthentic" self-disclosures. Specifically, we implemented a genetic algorithm that evolves prompts using various state-of-the-art prompt engineering techniques, including chain of thought, self-critique, generated knowledge, and expert prompting. Our results showed that, of the methods tested, the evolved self-critique prompt engineering method achieved the best result, with an F1-score of 0.8.

1 Introduction

Longitudinal changes in linguistic abilities have been studied to identify a relationship between language decline and the onset of dementia (Kempler and Goral, 2008). The Nun Study, a longitudinal investigation into Alzheimer's disease, examined this relationship (Kemper et al., 2001). Kemper et al. discovered in their study that higher linguistic abilities in early adulthood, measured by the proportion of complex sentences in writing samples, were linked to a lower risk of developing dementia. While longitudinal research offers valuable insights into causal relationships, it is often challenging and costly to collect such data (M. Leffler and Tong, 2022). Social media data has become a promising source for creating cohorts for longitudinal studies (Zubiaga, 2018), as data can be continuously and passively collected from users' interactions over extended periods. A further significant advantage of social media data is that each post is timestamped, making it easy to track changes over time. This allows researchers to analyze linguistic patterns with precise temporal context, capturing everyday language use across various contexts. This characteristic enhances the ability to study longitudinal changes in language and its relation to conditions such as dementia (Hrincu et al., 2022).

A key step in social media analysis, following the collection of user data, is the annotation process (Wongkoblap et al., 2022). Accurate annotation is vital, as correctly labelling users enables researchers to distinguish between groups and analyze their differences. While methods relying solely on pattern matching for the identification of self-disclosure statements are straightforward to implement, they often prove unreliable in the context of mental health and neurodegenerative condition due to the tendency of such disclosures to be humorous, whimsical, or sardonic.

In this research, we leverage Large Language Models (LLMs) to automate the annotation of social media data related to dementia self-disclosure. LLM performance is highly dependent on the quality of the prompts guiding the model. To optimize these prompts, we implemented a genetic algorithm that evolves them using various state-of-the-art (SOTA) prompt engineering techniques. By monitoring the performance of these techniques, we gained valuable insights into which methods are most effective for this task. Our prompts were also designed as detailed guidelines, enabling the model to detect subtle linguistic patterns critical to identifying authentic dementia-related disclosures. This approach not only improves annotation accuracy but also enhances interpretability, offering

Proceedings of The 22nd Annual Workshop of the Australasian Language Technology Association, pages 189–196
December 2-4, 2024 ©2024 Association for Computational Linguistics

researchers insights into the linguistic features of dementia self-disclosure on social media.

2 Related Work

2.1 Manual Annotation

A traditional approach to identifying users with health conditions involves manual annotation (Wongkoblap et al., 2022). In this method, a dataset is typically built by using keywords to scrape social network platforms, followed by manually annotating the collected data (Chancellor et al., 2023). For instance, Talbot et al. (2018) collected tweets containing search terms associated with Alzheimer's or dementia, such as "I have dementia," to identify users with self-reported diagnoses. While relying solely on search terms to label users as dementia patients is a simple way to annotate a dataset, it is prone to noise and incorrect labeling. For instance, the phrase "I have dementia" can appear in contexts that are not intended to be taken literally, such as jokes or memes—e.g., "My doctor said I have dementia. Well, I don't remember asking."

Similarly, Azizi et al. (2024) and Gkotsis et al. (2020) used the search terms "Dementia" or "Alzheimer" to collect data from Twitter and Reddit. However, in both studies, the collected data was manually filtered to remove irrelevant content where the search terms were not used to indicate that a person was suffering from these illnesses. This manual filtering process helped reduce noise, increasing the likelihood that posts genuinely related to dementia or Alzheimer's self-disclosure were retained for further analysis. While effective, this method still requires substantial human effort to ensure the accuracy of the annotations.

2.2 Automated Prompt Engineering

The performance of an LLM is tied to the quality of prompts used to instruct them. Chain-of-Thought (CoT) prompting encourages LLMs to incorporate intermediate reasoning steps, breaking down complex tasks into smaller, logical components (Wei et al., 2022). Generated Knowledge (GK) prompting augments the input with relevant information, effectively expanding the model's contextual understanding (Liu et al., 2022). Self-critique (SC) prompting introduces an additional layer of reflection, where the model is encouraged to assess and critique its own output (Wang et al., 2023). Expert prompting explicitly indicates to the LLM that it is proficient in a particular field; e.g. an expert in prompt engineering (Xu et al., 2023). Testing a diverse set of prompts is crucial for optimizing the output of an LLM, as it enables the model to explore a broader solution space and consider multiple approaches to a problem (Fernando et al., 2023).

Automated prompt strategies, aimed at minimizing manual intervention in prompt design and optimization, have demonstrated promising results (Cabrera Lozoya et al., 2024). In this paper, we leveraged LLMs to generate prompt candidates. We employed a binary tournament genetic algorithm framework (Harvey, 2009), which involves randomly selecting two prompts and replacing the prompt with lower fitness by a mutated version of the one with higher fitness.

3 Method

3.1 Data collection

To construct our dataset, we used the Twitter Academic API to collect tweets containing search terms like "I have dementia," yielding a total of 14,025 tweets. The data collection took place between October and November 2022. For each self-disclosure tweet, we also gathered the five posts immediately preceding and following the self-disclosure to assess their context. For the complete list of self-disclosure terms used for the data collection, please refer to Appendix A. Three authors of the paper were responsible for annotating the dataset. To improve inter-annotator agreement, they completed four annotation blocks, each consisting of 1,991 tweets. A substantial inter-annotator agreement was achieved, with a pairwise Cohen's kappa of 0.68 (McHugh, 2012). Of the tweets collected using the search terms, less than 20% were authentic. From the remaining data we built a balanced dataset with a 50/50 distribution of authentic and inauthentic statements by applying upsampling. The dataset was divided into stratified training and testing sets, following an 80/20 split. The training and testing datasets were verified to ensure there was no cross-contamination between them. The training dataset was then divided into 10 stratified batches.

3.2 Genetic Algorithm

Let P represent the prediction from an LLM when given an instruction prompt I as input, expressed as $P = \text{LLM}(I)$. Our genetic algorithm aims to find an optimal instruction prompt O with the goal of maximizing the performance of P in comparison

when I is utilized. Our algorithm mutates prompts to optimize them. Mutations involve a mutation prompt M and an LLM. A mutated prompt I' is defined as I' = LLM(M+ I), where + denotes string concatenation. The pool of mutation prompt types is derived from prompt engineering techniques employed to enhance prompts for LLMs. In our experiment we tested CoT, GK, SC, and Expert techniques. Appendix B contains the set of starting prompts for each type of mutation and a prompt mutation example.

Given an initial instruction prompt to label a tweet as originating from a user who authentically identifies themselves as having a diagnosis of dementia, our algorithm creates an initial population of prompts by evolving the initial instruction prompt using a set of random mutation prompts. The mutated prompts are then used by the LLM to make predictions on a random batch from the training dataset. Once the batch has been processed, the accuracy that the LLM obtained using each prompt is stored as the fitness level of that prompt. Our algorithm maintains a record of the instruction prompt, the mutation prompt, and the associated fitness level that the prompt achieved when processing a batch of tweets. Each record represents an individual in the population.

Once the population is initialized, our evolutionary process unfolds in generational steps. In each step, each individual has a mutation probability of μ_m, representing the likelihood of undergoing a mutation that alters its instruction prompt. After selecting which individuals will mutate, our algorithm then determines the type of mutation to be acquired from four options: CoT, GK, Expert, or SC. Upon calculating the mutated individual's fitness using a random batch from the training dataset, it is introduced into the population. This process continues until the maximum population cap is reached. Once the population cap is met, individuals for the next generation are selected using a probability function weighted by each individual's fitness level. This ensures that fitter individuals have a higher likelihood of advancing, while still allowing for some diversity by giving less fit individuals a chance to survive. After N generations, the instruction prompt from the individual with the highest fitness is selected as the optimized prompt. Figure 1 presents an overview of our algorithm.

3.3 Natural Language Processing Models

Our genetic algorithm was tested using Meta-Llama-3-8B-Instruct[1], with a nucleus sampling of 0.9 and a temperature of 0.6. Since the LLM can generate diverse textual outputs to label each tweet, we appended a formatting prompt instructing the model to respond with a 'yes' or 'no'. Subsequently, a BERT text classifier was utilized to categorize the LLM's outputs. A label of 0 indicated that the text did not come from a user who genuinely disclosed themselves as having dementia, while a label of 1 indicated the opposite, signifying genuine self-disclosure of a dementia diagnosis. This classification step ensures a standardized and consistent output, which was needed to measure the accuracy and F1 score of the LLM model. Refer to Appendix C for an example of a classification.

3.4 Evaluation

To find the optimal prompt, we executed the genetic algorithm with a population limit set to 10 individuals, a mutation probability μ_m of 50%, and spanning a total of 20 generations. Subsequently, we selected the prompt with the highest fitness level from the surviving population. The selected prompt became the input for the LLM, and we assessed its performance using the tweets from the testing dataset. Our evaluation metrics included measuring and reporting both the F1-score and the accuracy achieved by the LLM on the testing dataset. For comparison, we also trained and tested a BERT model, using it as a baseline to assess the performance of our algorithm against traditional transformer-based classifiers. Details of the BERT model's hyperparameters are presented in Appendix D.

4 Results and Discussion

The optimized prompt (refer to Appendix E) achieved an accuracy of 0.8 and an F1-score of 0.8, outperforming the BERT classifier, which obtained an accuracy of 0.7 and an F1-score of 0.71. In Figure 2, the distribution of mutation types among individuals across generations is illustrated.

The most prevalent mutation type observed throughout multiple generations stemmed from the SC prompt engineering technique, with the top-performing prompt from the final generation being a product of a SC mutation prompt. However,

[1]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

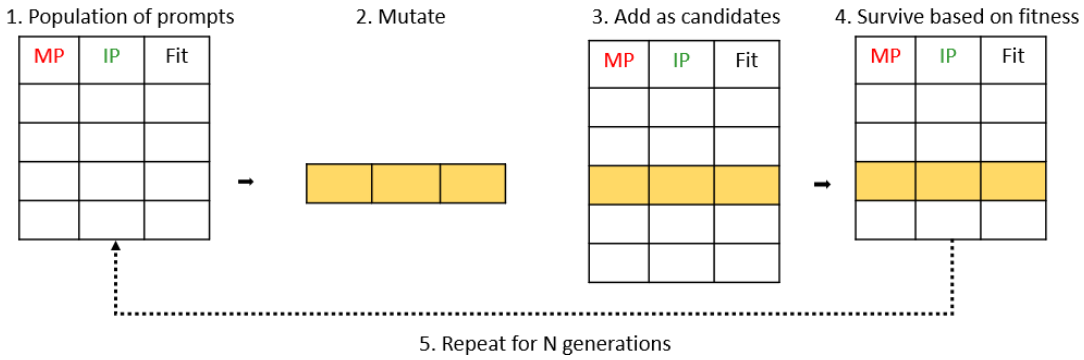

Figure 1: In our genetic algorithm, each individual has an instruction prompt (IP) guiding the LLM, a mutation prompt (MP) used to generate the instruction, and a fitness score based on the LLM's performance with that prompt. At each generational step, individuals have a probability of undergoing mutation, with the mutation type selected from a predefined pool. Mutated individuals are added to the population, and once the population cap is reached, a fitness-based probabilistic selection is applied to determine which individuals advance to the next generation.

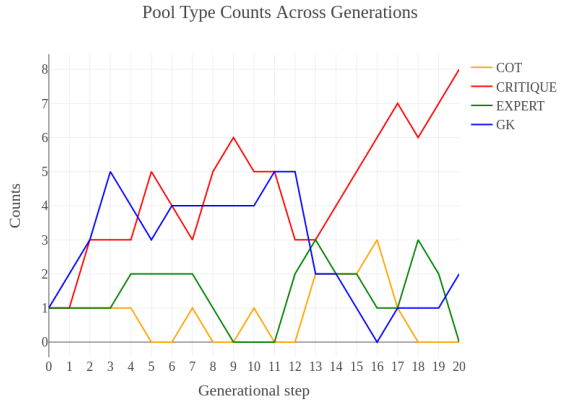

Figure 2: Number of individuals from a given mutation type in the population at a given generational step.

upon reviewing the prompt, we observed the integration of elements from various prompt engineering methods. Prompts derived from GK typically include an enumeration of components to evaluate. When followed by an SC mutation, the prompt addresses shortcomings in the components suggested and guides the model to contextualize them properly. Additionally, elements of CoT mutations are evident in the logical step-by-step structure of the prompt. All these characteristics were present in the optimized prompt. Therefore, our findings suggest that the optimal prompt engineering approach involves a blend of different techniques.

The adaptive prompt engineering technique was developed and evaluated using an open-access model that can be run locally, enabling researchers to analyze sensitive content without needing to send it to third-party organizations. Additionally, since the model is open-access, there are no associated usage fees, which reduces costs and improves accessibility, particularly in less well-resourced settings. Our algorithm also offers an accessible approach for public health researchers to identify self-diagnosed patients on social media for cohort building. It minimizes the need for expertise in machine learning or prompt engineering, as SOTA techniques are integrated into the algorithm. Moreover, our algorithm allows for upgrades upon the discovery of new prompt engineering techniques, requiring only their addition to the mutation pool.

5 Conclusion

We used a genetic algorithm to optimize prompts for LLMs to detect self-disclosed dementia statements in tweets. The optimal prompt achieved an accuracy of 0.8 and an F1 score of 0.8, surpassing the BERT classifier, which had an accuracy of 0.7 and an F1 score of 0.71. Additionally, it significantly outperformed a method that would solely rely on key search terms to label users as having dementia, as our annotation process revealed that less than 20% of the collected tweets with dementia self-disclosure statements were authentic. The algorithm used SOTA prompt engineering methods, and analysis revealed that SC mutations outperformed the other mutation types.

Although our algorithm was designed to automate the annotation of dementia-related data, it can also assist in the annotation of other types of data when provided with the appropriate datasets. We envision that by adapting our algorithm, researchers may find it helpful in supporting the annotation process across various domains, improving efficiency and reducing manual labor.

References

Mehrnoosh Azizi, Ali Akbar Jamali, and Raymond J Spiteri. 2024. Identifying X (formerly Twitter) posts relevant to dementia and Covid-19: Machine learning approach. *JMIR Formative Research*, 8:e49562.

Daniel Cabrera Lozoya, Jiahe Liu, Simon D'Alfonso, and Mike Conway. 2024. Optimizing multimodal large language models for detection of alcohol advertisements via adaptive prompting. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 514–525, Bangkok, Thailand. Association for Computational Linguistics.

Stevie Chancellor, Jessica L. Feuston, and Jayhyun Chang. 2023. Contextual gaps in machine learning for mental illness prediction: The case of diagnostic disclosures. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–27.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. https://arxiv.org/abs/2309.16797.

George Gkotsis, Christoph Mueller, Richard J.B. Dobson, Tim J.P. Hubbard, and Rina Dutta. 2020. Mining social media data to study the consequences of dementia diagnosis on caregivers and relatives. *Dementia and Geriatric Cognitive Disorders*, 49(3):295–302.

Inman Harvey. 2009. The microbial genetic algorithm. In *European Conference on Artificial Life*.

Viorica Hrincu, Zijian An, Kenneth Joseph, Yu Fei Jiang, and Julie M. Robillard. 2022. Dementia research on Facebook and Twitter: Current practice and challenges. *Journal of Alzheimer's Disease*, 90(2):447–459.

Susan Kemper, Lydia H. Greiner, Janet G. Marquis, Katherine Prenovost, and Tracy L. Mitzner. 2001. Language decline across the life span: Findings from the nun study. *Psychology and Aging*, 16(2):227–239.

Daniel Kempler and Mira Goral. 2008. Language and dementia: Neuropsychological aspects. *Annual Review of Applied Linguistics*, 28:73–90.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Grace M. Leffler and Xin Tong. 2022. A tutorial on collecting and processing longitudinal social media data. *International Journal of Arts, Humanities amp; Social Science*, 03(10):21–29.

Marry L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, page 276–282.

Catherine Talbot, Siobhan O'Dwyer, Linda Clare, Janet Heaton, and Joel Anderson. 2018. Identifying people with dementia on Twitter. *Dementia*, 19(4):965–974.

Rui Wang, Hongru Wang, Fei Mi, Yi Chen, Ruifeng Xu, and Kam-Fai Wong. 2023. Self-critique prompting with large language models for inductive instructions. https://arxiv.org/abs/2305.13733.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Akkapon Wongkoblap, Miguel A. Vadillo, and Vasa Curcin. 2022. *Social media big data analysis for mental health research*, page 109–143. Elsevier.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. https://arxiv.org/abs/2305.14688.

Arkaitz Zubiaga. 2018. A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8):974–984.

A Search terms

Table 1 presents the terms used to search for individuals that self diagnosed with dementia.

B Mutation example

Table 2. presents the initial prompts for each type of mutation. Figure 3 illustrates an example of a mutation step. In this scenario a mutated prompt is created by using a mutation prompt from the Generated Knowledge (GK) pool to mutate an instruction prompt.

C Classification example

Figure 4 illustrates the classification process used in the genetic algorithm.

D BERT hyperparameters

The BERT model was trained for a total of 3 epochs, using an AdamW optimizer, with an initial learning rate of 5×10^{-5}, and a weight decay of 0.01.

E Optimized prompt

The optimal prompt after 20 generational steps was:

Determine whether the following text is from a user genuinely self-identifying as having Dementia or Alzheimer's by analyzing the language, syntax, and content for the following criteria:

- *Presence of medical terms such as 'Dementia', 'Alzheimer's','memory loss', 'cognitive decline', or 'neurodegenerative disease', and assess their usage in context to avoid mere knowledge-based mentions, considering the user's level of medical knowledge and awareness, as well as the specificity and relevance of the terms to their personal experience.*
- *Use of first-person narrative, indicating a personal experience with the condition, including specific details about the onset, progression, or impact on daily life, and evaluate the coherence, consistency, and emotional authenticity of the narrative.*
- *Expression of emotional distress, such as anxiety, fear, or frustration related to the condition, and evaluate the intensity, authenticity, and emotional regulation of the user's language, considering factors such as the use of emotional language, personal anecdotes, and vulnerability.*
- *Disclosure of symptoms, including memory lapses, confusion, difficulty with speech, or changes in mood, and consider the specificity, consistency, and progression of the symptom descriptions over time, as well as the user's level of awareness and understanding of their symptoms.*

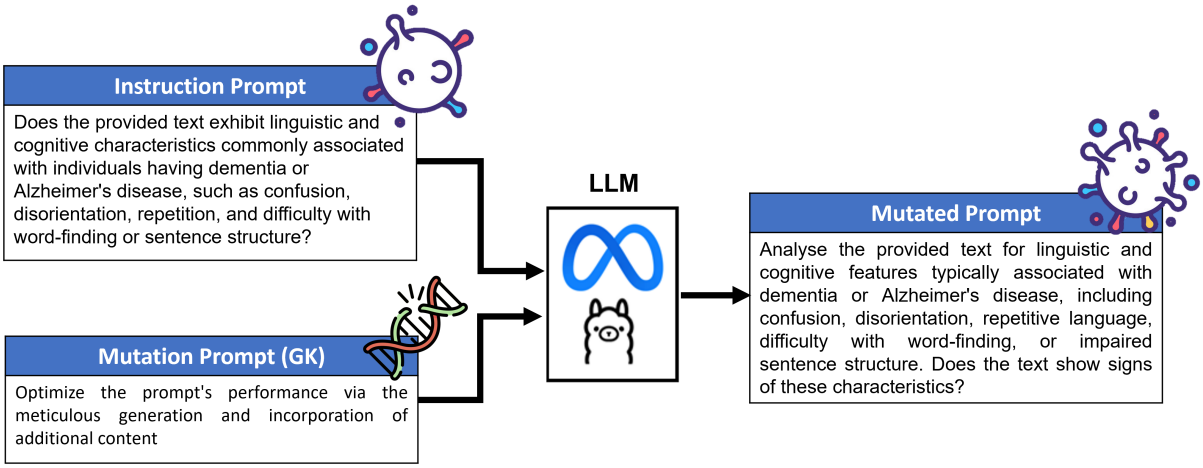

Figure 3: Example of a mutation step.

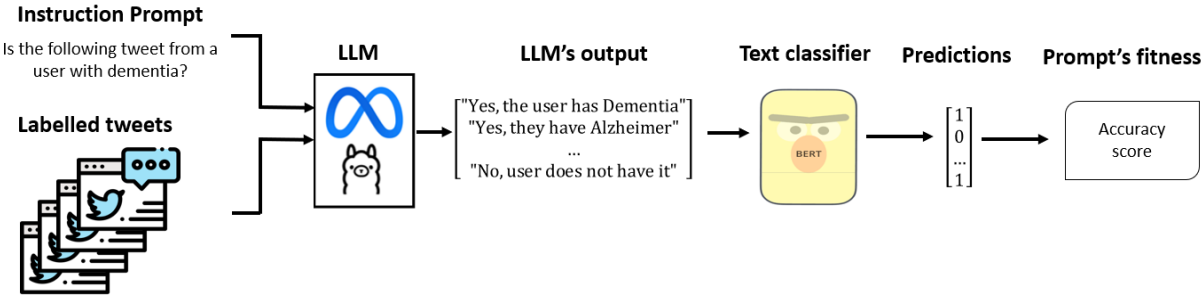

Figure 4: Example of a mutation step.

Dementia search terms	
I have lewy body	I have dementia with lewy bodies
I was diagnosed with lewy body	I was diagnosed with dementia with lewy bodies
I've been diagnosed with lewy body	I've been diagnosed with dementia with lewy bodies
I've got lewy body	I've got dementia with lewy bodies
Just been diagnosed with lewy body	Just been diagnosed with dementia with lewy bodies
I have dementia	I've been diagnosed with dementia
I've got dementia	Just been diagnosed with dementia
I have vascular dementia	I was diagnosed with vascular dementia
I've been diagnosed with vascular dementia	I've got vascular dementia
Just been diagnosed with vascular dementia	I have alzheimers
I was diagnosed with alzheimers	I've been diagnosed with alzheimers
I've got alzheimers	Just been diagnosed with alzheimers

Table 1: Search terms used to collect self-disclosure statements from Twitter.

Mutation type	Prompts
Chain of thought	Append to the following instruction the following text, "Let's think step by step."
	Decompose and rewrite the instruction as a set of logical steps, rewrite it as a sentence.
	Rewrite the following instruction by adding intermediate steps to enhance its performance.
Expert	Act as an expert in prompt engineering with 10 years of experience designing and debugging prompts. Identify the strengths and weaknesses of the following instruction, think about what changes you would make, and suggest an improved version.
	Imagine you are an expert in generating instructions for large multimodal models. You are designing an instruction to achieve the best possible result. A colleague shares their best instruction with you; identify why it is good and generate an even better one.
	Simulate being an expert program in improving instructions, detecting their strengths, weaknesses, and consistently providing better results. Take this prompt and make it better.
Generated Knowledge	Enhance the effectiveness of the following prompt by generating and appending additional content. Focus on providing specific examples, detailed criteria, or relevant guidelines to elevate its performance.
	Improve the prompt's performance through the strategic generation and integration of supplementary content, fostering heightened efficacy within the experimental domain.
	Optimize the prompt's performance via the meticulous generation and incorporation of additional content.
Critique	Critique the following instruction and propose enhancements to address any identified shortcomings. Please provide only the refined version in your response.
	Review the given instruction, identify any areas for improvement, and suggest changes to enhance its quality. Please provide a refined version that incorporate these improvements.
	Examine the given instruction, analyze it for potential shortcomings, and suggest improvements to address any identified issues. Submit only the refined version in your response, integrating enhancements to elevate its overall quality.

Table 2: Starting prompts for each mutation type.