

BingAster at #SMM4H-HeaRD 2025: Identifying Dementia Caregivers on Twitter Using Prompt-Based LLMs and Cognitive Distortion Patterns

Artin Tonekaboni,¹ Xin Wang,¹ Sadamori Kojaku,¹ Luis M. Rocha^{1,2}

¹ Binghamton University, New York, USA

² Universidade Católica Portuguesa, Lisbon, Portugal

atonekaboni@binghamton.edu, xwang314@binghamton.edu, skojaku@binghamton.edu, rocha@binghamton.edu

Abstract

Family caregivers of individuals with dementia often experience significant emotional and cognitive burden, which may manifest in the language they use on social media. Identifying such posts is valuable for understanding caregiver needs and advancing mental health research. Task 3 of the SMM4H 2025 shared task focuses on classifying whether a tweet indicates that the user has a family member with dementia. The task requires sensitivity to both direct and indirect expressions of caregiving. We addressed this task using a prompt-based zero-shot classification system powered by large language models (LLMs). Our method leverages instruction-tuned models, including DeepSeek-R1 and Mixtral-8x7B. To further evaluate our results, we developed a LLM-based multi-agent system to analyze cognitive distortions in tweets labeled as caregiver-related. The resulting distortion patterns offer psychological insight into the model’s predictions and highlight the system’s potential for broader applications in mental health monitoring.

Introduction

Family caregivers of people with dementia often face substantial emotional, physical, and financial challenges. While traditional in-person interventions, such as support groups and educational workshops, have shown benefits in reducing stress and improving coping skills, they often face logistical barriers, including time, location, and caregiving demands (Wen et al. 2023). Online interventions have emerged as a promising option, offering greater accessibility and convenience. Evidence suggests that internet-based interventions can reduce caregiver burden, depressive symptoms, and stress while improving self-efficacy (Leng et al. 2020; de Moraes-Ribeiro et al. 2024), though challenges, such as digital literacy gaps, engagement issues, and high dropout rates, remain (Wen et al. 2023).

To address these limitations, internet-based interventions have evolved to include a wider variety of digital formats and platforms. Social media platforms like Facebook offer peer support and reduce caregiver isolation (Wilkerson et al. 2018). Beyond social networking, artificial intelligence (AI) technologies provide personalized support and real-time responses through chatbots and recommendation sys-

tems (Leng et al. 2020). Though still emerging, AI tools show potential to scale and enhance caregiving support, but further research is needed to assess their long-term impact and ethical considerations (Baruah et al. 2021).

Task Description

This study supports the use of Twitter for internet-based interventions aimed at family caregivers of people with dementia. It focuses on Task 3 of the SMM4H 2025 shared task, which aims to identify whether a tweet suggests that the user has a family member with dementia. This challenge requires models to detect subtle, indirect language and understand emotional context, which is essential for identifying caregiving narratives in real-world social media posts.

The dataset includes binary-labeled tweets. A label of 1 indicates that the tweet suggests the user has a family member with dementia, while a label of 0 denotes no such indication. The dataset consists of 6,724 training tweets, 353 validation tweets, and 10,000 unlabeled test tweets. Each tweet includes an ID, raw text, and a label, with no text preprocessing to preserve informal language features.

Methodology

Prompt-based classification

This task focuses on a binary tweet classification problem, aiming to automatically identify the English-language tweets for those with a family member with dementia. To address this classification task, we employed a prompt-based approach using LLMs. This method involves presenting the model with structured natural language instructions and interpreting its output as a binary classification. We adopted a zero-shot learning setting, using the LLM without any fine-tuning process. This allows us to isolate the influence of prompt design and assess the model’s capacity to infer caregiver context from subtle, often indirect expressions in user-generated text.

Traditional approaches such as transformer-based model or support vector machine (SVM) are effective for well-defined text tasks, but they often underperform on informal, ambiguous, or emotionally nuanced language found in social media posts. We instead opted for decoder-only LLMs, which have demonstrated strong reasoning capabilities when guided with structured prompts.

Prompt-based classification is particularly well-suited for this task because it allows zero-shot inference without any feature engineering or task-specific retraining, while still capturing nuanced caregiving language. Unlike encoder-based models, decoder-style LLMs can generate responses that incorporate intermediate reasoning steps and structured output formats. Recent studies have shown that large language models, when guided by well-crafted prompts, can perform competitively on complex health-related classification tasks without additional training data (Zhang et al. 2024; Sharma, Chen, and Resnik 2024).

We designed a structured prompt to guide the large language model in performing the binary classification task, as presented in Appendix. The prompt consists of three main components: (1) a clear instruction defining the task, which determines whether the tweet indicates the user has a family member with dementia; (2) a reasoning cue—“Please think step by step”—which encourages the model to engage in a more deliberate reasoning process before producing a final output; and (3) an explicit labeling instruction, which standardizes the format of the model’s response to 1 or 0.

This output structure—“LABEL: 0” or “LABEL: 1”—was chosen to allow deterministic post-processing using simple regular expressions. This avoids ambiguity in response parsing and improves system reliability.

The inclusion of the phrase “think step by step” is inspired by chain-of-thought prompting techniques, which have been shown to improve model performance on reasoning tasks by encouraging intermediate reasoning steps before a final decision (Wei et al. 2022). This helps the model navigate ambiguous or implicitly phrased tweets more effectively.

LLM setup for classification task

We evaluated two large language models: **Mixtral-8x7B Instruct** and **DeepSeek-R1:70B**. Mixtral-8x7B Instruct is an open-weight, instruction-tuned Mixture-of-Experts (MoE) model released by Mistral AI. It consists of 8 experts with 7 billion parameters each, with 2 experts activated per forward pass. This sparse routing mechanism allows for efficient inference while preserving performance in reasoning tasks. DeepSeek-R1:70B is a large-scale bilingual model trained on both English and Chinese corpora, derived from the LLaMA architecture. It was optimized for instruction following and general-purpose classification tasks using reinforcement learning rather than traditional supervised fine-tuning.

Both models were hosted and queried locally using the Ollama framework for local inference. Temperature was fixed at 0.0 to ensure deterministic responses. Prompts followed a standardized instruction format, and labels were extracted using regular expressions that matched outputs of the form “LABEL: 0” or “LABEL: 1”.

Results

We evaluated model performance using standard classification metrics: precision, recall, and F1-score. All experiments were conducted on raw tweet texts, without any pre-processing, feature engineering, or model fine-tuning. This

setup was intended to isolate the effectiveness of prompt-based zero-shot classification. Table 1 summarizes the results obtained on the validation set by two large language models: Mixtral-8x-7B and DeepSeek-R1:70B.

| Model | Precision | Recall | F1 Score |
|-----------------|-----------|--------|----------|
| Mixtral-8x-7B | 0.8511 | 0.8547 | 0.8529 |
| DeepSeek-R1:70B | 0.9035 | 1.000 | 0.9493 |

Table 1: Performance on validation set using Mixtral-8x-7B and DeepSeek-R1:70B.

Among the models evaluated, DeepSeek-R1:70B consistently outperformed Mixtral-8x-7B across all metrics. Most notably, DeepSeek achieved perfect recall (1.000), indicating its ability to identify nearly all caregiver-related tweets. This suggests superior sensitivity to subtle or implicit expressions of caregiving roles, such as tweets like “I miss how my grandma used to be”, which were often missed by Mixtral-8x-7B. The higher F1-score also reflects better balance between precision and recall, demonstrating DeepSeek’s overall robustness.

To further validate the model performance, we used the DeepSeek-based prompt system to perform evaluation. Table 2 shows the results.

| Dataset | F1 Score | Precision | Recall |
|----------|----------|-----------|--------|
| Test Set | 0.942 | 0.899 | 0.991 |

Table 2: Official test set performance for our DeepSeek-based system.

Compared to other SMM4H 2025 submissions (mean F1: 0.885; median F1: 0.953), our system achieved a great F1 score, indicating strong overall effectiveness. In particular, it excelled in recall (0.991), significantly outperforming the average, and demonstrating superior sensitivity to a wide range of caregiver-related language.

We hypothesize that DeepSeek-R1:70B outperformed Mixtral-8x7B due to architectural and training differences. DeepSeek is a dense model with 70 billion parameters, all of which are active during inference, allowing it to model more subtle semantic cues. It is instruction-tuned and further optimized via reinforcement learning from human feedback (RLHF), improving its reasoning and classification reliability (AI 2024). In contrast, Mixtral-8x7B is a Mixture-of-Experts (MoE) model in which only two of eight experts are active per inference, effectively reducing the model capacity at runtime (AI 2023). While this improves efficiency, it may limit sensitivity to indirect or emotionally nuanced language, leading to slightly lower performance on our task.

Evaluation by Cognitive Distortion Analysis

To further assess the reliability of our classification system beyond standard performance metrics, we conducted a post-hoc analysis of cognitive distortions inferred tweets labeled as caregiver-related. Prior research suggests that dementia caregivers frequently develop cognitive distortions—recurring, biased thought patterns such as *catastrophism*

or *emotional reasoning*—which can ultimately affect their emotional well-being negatively (Losada et al. 2006). For robustness of evaluation, we compared cognitive distortion between sets to assess the consistency of psychological signals.

Cognitive distortion analysis identifies irrational or exaggerated thought patterns commonly associated with, or even driving, anxiety, depression, or caregiver burnout (Beck 1979; Burns 1980). In natural language analysis it has been shown to reveal implicit cognitive dispositions—such as overgeneralizing negative outcomes or discounting positive experiences—that may not be captured through sentiment or topic-based classification alone (Bollen et al. 2021).

To comprehensively analyze cognitive distortions in tweets, we built a multi-agent system with DeepSeek-R1:32B to instantiate a collective decision-making process. In this system, a controller agent distributes each tweet’s text to 11 agents. Each such agent was tasked with evaluating if a unique cognitive distortion dimension is present in a tweet, based on the definition of that dimension and its associated set of linguistic features obtained from an established cognitive distortion corpus (Bollen et al. 2021). After each distortion-specific agent issues an evaluation, the controller aggregates the binary outputs into a final multi-label distortion vector. In Appendix prompts for this multi-agent system are provided, and Table 3 lists representative linguistic cues for the 11 cognitive distortion types.

The distribution of cognitive distortion dimensions per dataset splits are shown in Figure 1. Tweets labeled as positive in the training and validation sets—where ground truth annotations are available—exhibited consistently elevated distortions in dimensions such as emotional reasoning, catastrophizing, and personalizing. These patterns generally align with previously observed psychological observations of caregiver stress (Losada et al. 2006), though further analysis is needed to characterize elevated dimensions such as mental filtering, which, to the best of our knowledge, have not been previously reported for caregivers. Importantly, the training and validation (positive-label) tweet sets show a similar distortion distribution. This suggests that our classifier is reliably identifying caregiver language by capturing not just surface-level patterns, but deeper psychological indicators associated with caregiving burden. This post-hoc analysis supports the LLM classification predictions, provides additional validation of classifier generalization, and provides psychological explainability of positive-label tweets.

Conclusion

In this study, we presented a prompt-based classification system that leverages LLMs to identify whether a tweet indicates the user has a family member with dementia. Our approach applied instruction-tuned models with zero-shot prompting, instead of involving model fine-tuning and handcrafted features. Evaluation on both validation and test demonstrated strong performance of the proposed system. To further assess its reliability, we conducted a cognitive distortion analysis on tweets authored by caregivers. By comparing distortion patterns across training, validation, and

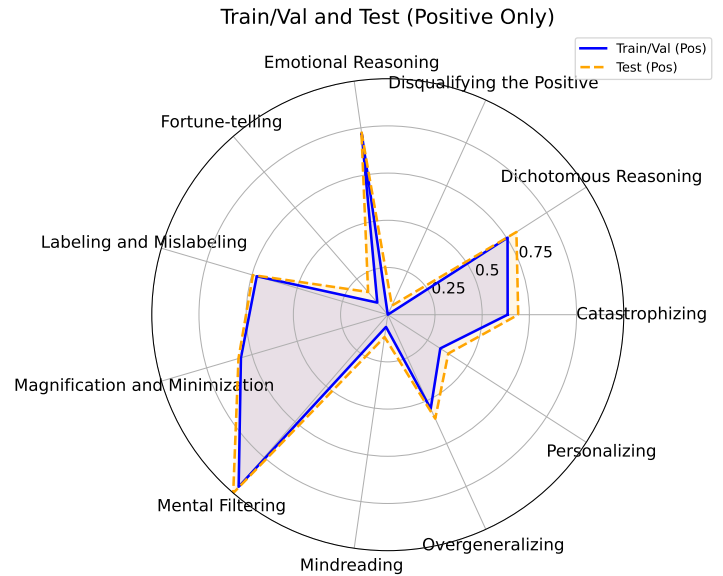


Figure 1: Distribution of cognitive distortion dimensions across all dataset splits. Positive-labeled tweets consistently show elevated levels of distortions such as emotional reasoning and catastrophizing, aligning with the known psychological stress experienced by caregivers (Losada et al. 2006).

test sets, we observed that tweets predicted as caregiver-authored exhibited psychological signals consistent with known caregiver stress behavior. These findings suggest that our model effectively generalizes to unseen data while preserving meaningful psychological patterns.

References

AI, D. 2024. DeepSeek R1 Models. <https://github.com/deepseek-ai>. Accessed: 2025-05-06.

AI, M. 2023. Mixtral of Experts. <https://mistral.ai/news/mixtral-of-experts/>. Accessed: 2025-05-06.

Baruah, U.; Varghese, M.; Loganathan, S.; Mehta, K. M.; Gallagher-Thompson, D.; Zandi, D.; Dua, T.; and Pot, A. M. 2021. Feasibility and preliminary effectiveness of an online training and support program for caregivers of people with dementia in India: a randomized controlled trial. *International journal of geriatric psychiatry*, 36(4): 606–617.

Beck, A. T. 1979. *Cognitive therapy and the emotional disorders*. Penguin.

Bollen, J.; Ten Thij, M.; Breithaupt, F.; Barron, A. T.; Rutter, L. A.; Lorenzo-Luaces, L.; and Scheffer, M. 2021. Historical language records reveal a surge of cognitive distortions in recent decades. *Proceedings of the National Academy of Sciences*, 118(30): e2102061118.

Burns, D. 1980. *Feeling good: The new mood therapy*. William Morrow and Company, Inc., New York, New York.

de Moraes-Ribeiro, F. E.; Moreno-Cámara, S.; da Silva-Domingues, H.; Palomino-Moral, P. Á.; and del Pino-Casado, R. 2024. Effectiveness of Internet-Based or Mobile App Interventions for Family Caregivers of Older Adults with Dementia: A Systematic Review. In *Healthcare*, volume 12, 1494. MDPI.

Leng, M.; Zhao, Y.; Xiao, H.; Li, C.; Wang, Z.; et al. 2020. Internet-based supportive interventions for family caregivers of people with dementia: systematic review and meta-analysis. *Journal of medical Internet research*, 22(9): e19468.

Losada, A.; Montorio, I.; Knight, B. G.; Márquez, M.; and Izal, M. 2006. Explanation of caregivers distress from the cognitive model: the role of dysfunctional thoughts. *Psicología Conductual*, 14(1): 115.

Sharma, I.; Chen, E.; and Resnik, P. 2024. Mental Health Disclosure Detection in Social Media Using Prompt-Based Language Models. *SAGE Open*, 14(1): 1–14.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wen, Y.; Xing, Y.; Ding, Y.; Xu, W.; and Wang, X. 2023. Challenges of conducting of online educational programs for family caregivers of people with dementia living at home: An integrative review. *International Journal of Nursing Sciences*, 10(1): 121–128.

Wilkerson, D. A.; Brady, E.; Yi, E.-H.; and Bateman, D. R. 2018. Friendsourcing peer support for Alzheimer’s caregivers using Facebook social media. *Journal of Technology in Human Services*, 36(2-3): 105–124.

Zhang, R.; Lee, J. D.; Sun, C.; Goharian, N.; and Friedenberg, F. 2024. Social Media Mining for Pharmacovigilance Using Large Language Models: A Case Study on Adverse Drug Reactions. *Journal of Biomedical Informatics*, 145: 104454.

Definition: [DEFINITION]

Linguistic features: [FEATURE LIST].

Linguistic feature samples on cognitive distortion

| Cognitive distortion | Linguistic feature samples |
|-----------------------------|---|
| Catastrophizing | will fail; everything will collapse |
| Dichotomous Thinking | always; never; all or nothing |
| Disqualifying the Positive | good but not enough; that doesn’t count |
| Emotional Reasoning | I feel like a failure; because I feel it, it’s true |
| Fortune-Telling | I will mess up; it’s going to go badly |
| Labeling | I’m worthless; she’s a loser |
| Magnification/ Minimization | this is the worst; it’s not a big deal |
| Mental Filtering | only seeing what’s wrong; ignoring successes |
| Overgeneralization | nothing ever works; everyone leaves me |
| Personalization | it’s my fault; I caused this |
| Should Statements | I should have known better; I must succeed |

Table 3: Linguistic feature samples of 11 cognitive distortions.

Appendix

LLM prompt for classification task

Identify if the person who posted the following Twitter tweet has a family member with dementia. Please think step by step and then give the label result. If this person has a family member with dementia, reply "LABEL: 1". Otherwise, reply "LABEL: 0". The Twitter tweet is: [TEXT]

System prompt for multi-agent system

Does this sentence reflect [COGNITIVE DISTORTION NAME] based on the following definition and linguistic features? Think step by step and return 0 or 1.