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]
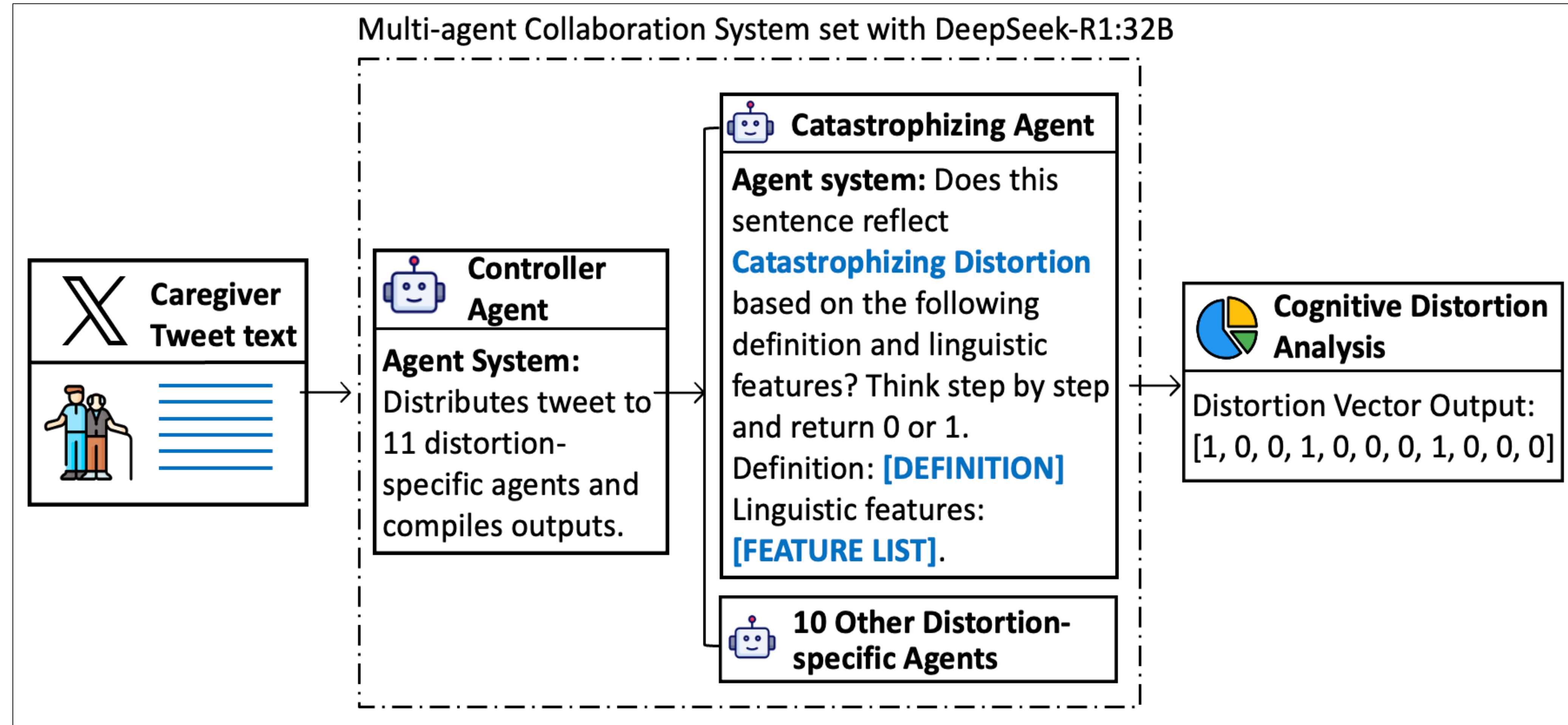[1]Binghamton University, [2]Universidade Católica Portuguesa
Corresponding author email: xwang314@Binghamton.edu
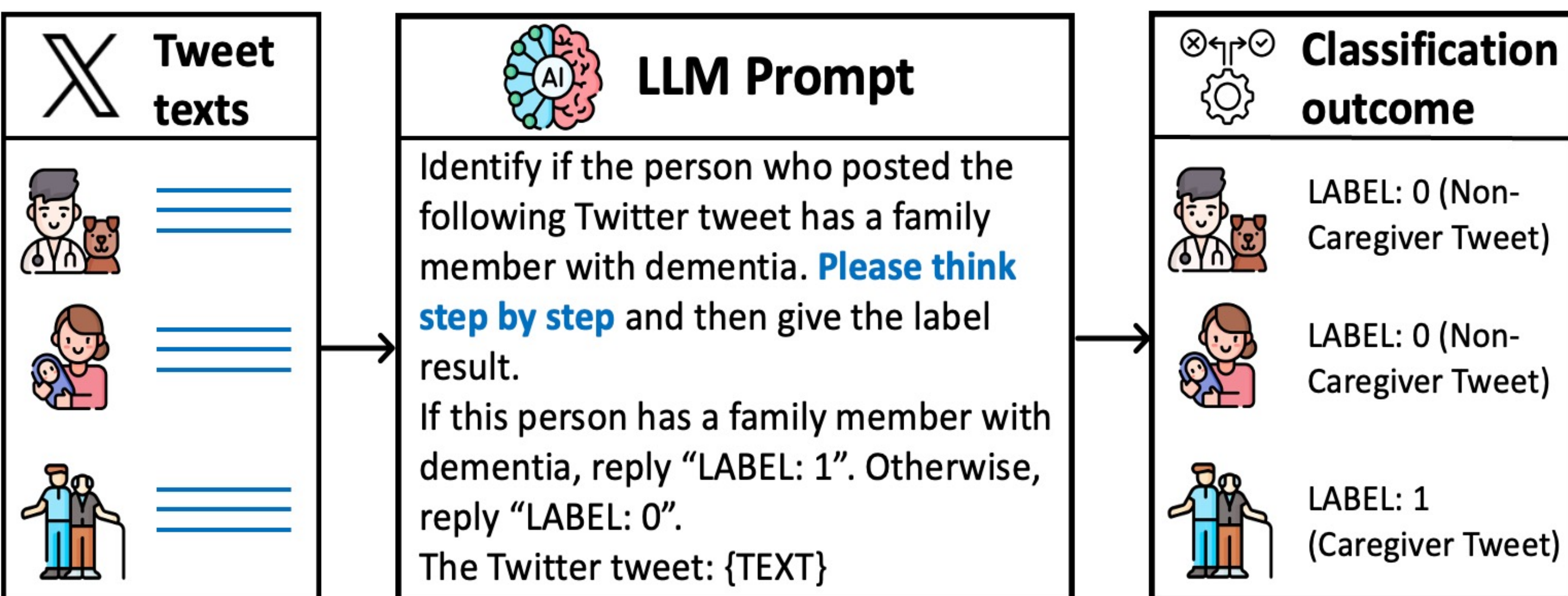
Project page

Introduction

- Family members looking after someone with dementia routinely shoulder substantial emotional, physical, and financial strain, which often reflects in the language they use online.
- **Our study tackles this identification challenge.** For SMM4H-2025 Task 3, we develop a prompt-based, zero-shot large-language-model system that flags tweets indicating a family member with dementia, laying groundwork for scalable, AI-driven caregiver support.
- **We developed a multi-agent system to analyze cognitive distortions** in detected caregiver tweets, helping further validate our model by uncovering stress-related thinking patterns like catastrophizing and emotional reasoning.
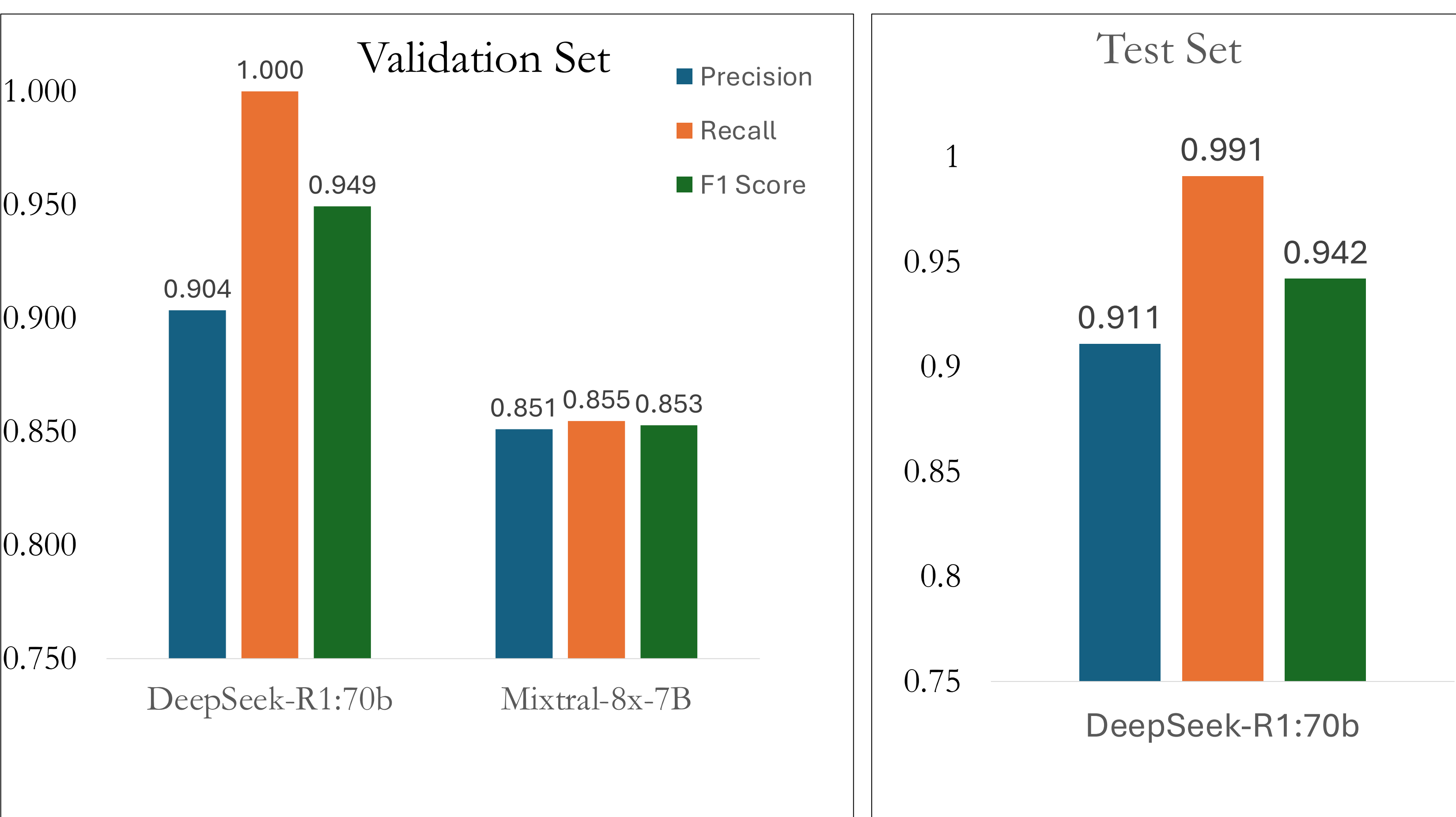
Multi-agent System for Cognitive Distortion Analysis

Multi-agent Collaboration System set with DeepSeek-R1:32B

Caregiver Tweet text → **Controller Agent**

Agent System: Distributes tweet to 11 distortion-specific agents and compiles outputs.

Catastrophizing Agent

Agent system: Does this sentence reflect **Catastrophizing Distortion** based on the following definition and linguistic features? Think step by step and return 0 or 1.
Definition: **[DEFINITION]**
Linguistic features: **[FEATURE LIST]**.

10 Other Distortion-specific Agents

Cognitive Distortion Analysis

Distortion Vector Output:
[1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0]

Prompt-based Classification

Tweet texts → **LLM Prompt** → **Classification outcome**

LLM Prompt: 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: {TEXT}

Classification outcome:
LABEL: 0 (Non-Caregiver Tweet)
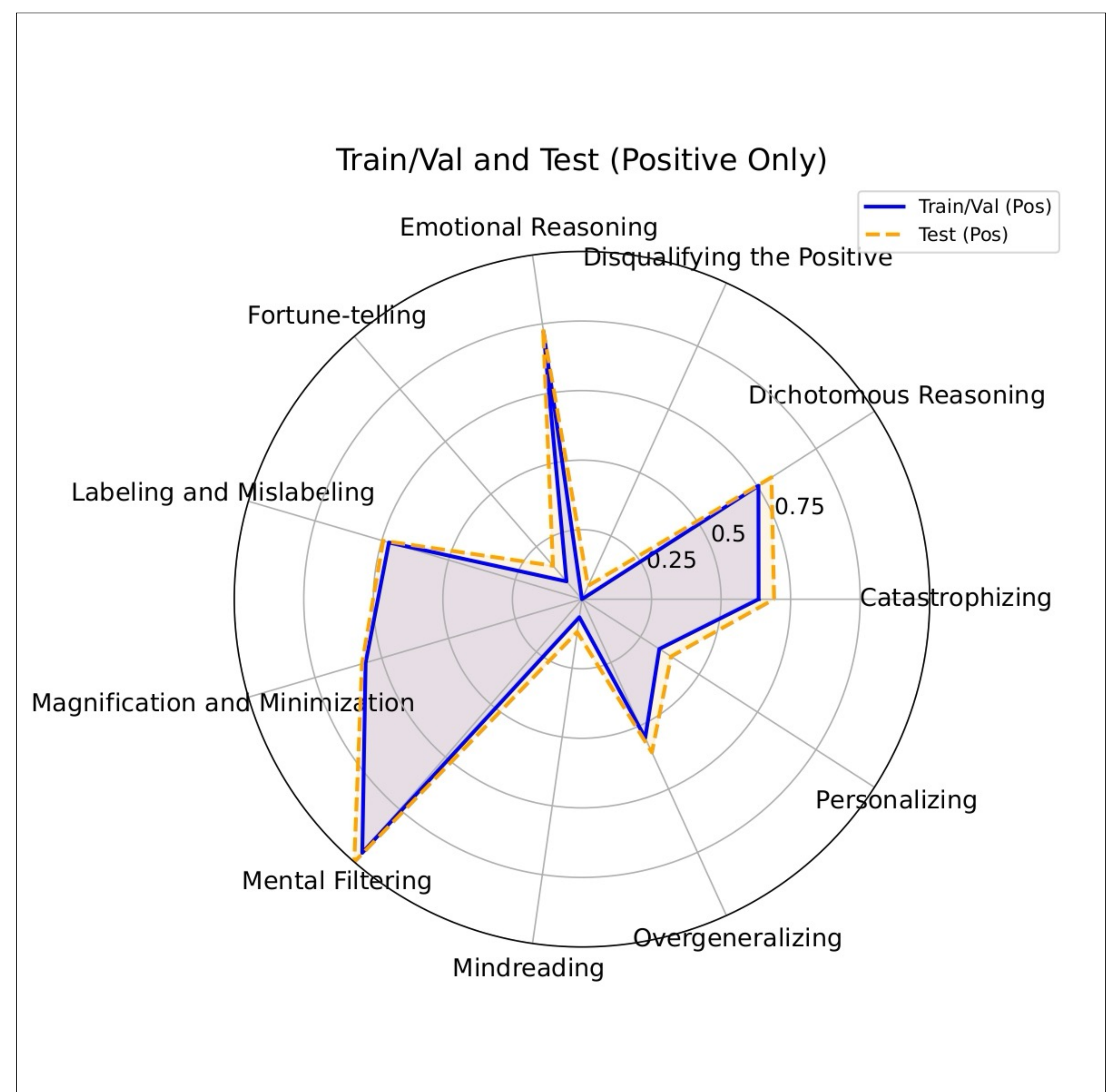LABEL: 0 (Non-Caregiver Tweet)
LABEL: 1 (Caregiver Tweet)

- A zero-shot prompt system is designed to classify tweets as written by a caregiver or not.
- "think step by step" can trigger LLM to conduct chain-of-thought reasoning in this binary classification task.

Results

Validation Set — Precision, Recall, F1 Score
- DeepSeek-R1:70b: 0.904, 1.000, 0.949
- Mixtral-8x-7B: 0.851, 0.855, 0.853

Test Set — DeepSeek-R1:70b: 0.911, 0.991, 0.942

- DeepSeek-V1.3 achieved the top F1 score (0.904) on the validation set, outperforming other approaches. It also performed well on the test set (F1 = 0.623), showing strong generalization.

Evaluation by Cognitive Distortion Analysis

Train/Val and Test (Positive Only)
— Train/Val (Pos), Test (Pos)

Axes: Emotional Reasoning, Disqualifying the Positive, Dichotomous Reasoning, Catastrophizing, Personalizing, Overgeneralizing, Mindreading, Mental Filtering, Magnification and Minimization, Labeling and Mislabeling, Fortune-telling

- The radar chart shows normalized distortion frequencies for test set (orange) and train/validation set (blue) tweets. Emotional Reasoning, Catastrophizing, and Mental Filtering appear more frequently in caregiver posts.
- These patterns reflect elevated emotional and cognitive stress in caregiver tweets, aligning with known psychological traits. Their consistency across data splits also provides indirect validation that the model is accurately identifying caregivers based on meaningful linguistic signals.

Conclusion and Future Work

- The distortion patterns across training, validation, and test sets suggest that our model effectively generalizes to unseen data while preserving meaningful psychological patterns.
- Future work will focus on improving classification accuracy, expanding cognitive distortion analysis, and designing adaptive online interventions to support caregivers in real time.