Cognitive Distortion Detection Using Large Language Models

Our research on cognitive distortion detection builds upon several key works in the field. We began by analyzing Shreevastava and Foltz's (2021) paper which established initial baselines using BERT-based models. Their work achieved F1 scores of 0.74 with logistic regression and 0.79 with SVM for binary classification, though they encountered limitations in multiclass settings with macro F1 score below 0.3. Understanding these limitations helped shape our approach to improving classification performance.

To ensure our implementation was clinically grounded, we dove deep into CBT literature. Burns' (1981) "Feeling Good: The New Mood Therapy" provided precise definitions of cognitive distortions that we could operationalize in our prompts. Additionally, Judith S. Beck's (2020) "Cognitive Behavior Therapy" offered structured frameworks for identifying distortion patterns. These clinical resources proved invaluable in designing prompts that could effectively guide LLMs in detecting cognitive distortions.

Based on this foundation, we developed enhanced prompting strategies that incorporated clinical expertise from CBT literature. Our implementation using state-of-the-art LLMs like Llama 3.1:8b and Llama 3.2:3b achieved substantial improvements over previous baselines. In binary classification, we reached an F1 score of 0.8124 with Llama 3.1:8b one-shot learning, while also improving macro F1 scores across all distortion types in multiclass classification.

Our work demonstrates how combining careful study of previous implementations with deep understanding of clinical frameworks can advance the state-of-the-art in cognitive distortion detection. The improvements we achieved suggest promising directions for future development of automated mental health support tools.

Research Context

- Primary Focus: Automated detection of cognitive distortions in therapeutic conversations-
- Dataset:
    - Utilized Cognitive Distortion Detection Dataset (2,568 classified statements)
    - Created 80/20 train/test split (2,062 training/506 test samples)
    - Covered 11 distinct categories of cognitive distortions

Prompting Frameworks

1. Baseline Framework

The basic zero-shot framework provides a foundation for cognitive distortion detection without requiring training examples. This approach relies on clear definitions of cognitive distortions and direct classification.

Components:

- Basic definition prompt: Presents clear, concise definitions of each cognitive distortion
- Enhanced definition variant: Includes more detailed descriptions and specific markers
- Direct classification approach without intermediate steps
- No example-based learning required

2. Conditional Framework

The conditional framework implements a structured, criteria-based approach to distortion detection, utilizing a two-level verification system to improve accuracy and reliability.

Components:

- Primary criteria (MUST have): Essential characteristics that define each distortion

- Secondary criteria (Supporting): Additional elements that confirm the presence of a distortion
- Systematic verification process
- Structured decision-making protocol

3. Expert-Based Framework
This framework leverages the concept of expert knowledge by positioning the LLM as a CBT specialist, incorporating clinical expertise into the detection process.

Variants:
1. Expert-Explanation: Combines expert positioning with detailed reasoning
2. Expert-Only: Focuses on expert decision-making without explanation
3. Explanation-Only: Emphasizes reasoning process without expert positioning

4. Hierarchical Framework
The hierarchical framework breaks down the detection process into distinct stages, allowing for more precise and systematic classification.

Implementations:
1. Hierarchy Baseline
    - Two-stage classification process
    - Sequential decision-making
    - Progressive refinement
2. Hierarchy-Expert
    - Incorporates expert perspective
    - Maintains clinical relevance
    - Enhanced decision precision
3. Hierarchy-Expert-Explanation
    - Combines expert positioning
    - Includes detailed reasoning
    - Comprehensive analysis process
4. Hierarchy-Explanation
    - Focus on reasoning process
    - Staged decision-making
    - Explicit analysis requirements

5. LLM-Defined Framework
This approach uses multi-turn prompting with Gemma to generate and refine cognitive distortion definitions, creating a more nuanced and AI-native understanding of distortion patterns. This framework represents an approach where we leverage one LLM (Gemma) to create definitions that other LLMs can better understand and apply, creating a more robust and AI-native detection system.

6. Elimination-Based Prompting:
- Developed systematic multi-turn dialogue system
- Progressive elimination of unlikely distortions
- Contextual maintenance across dialogue turns
- Final refinement to specific distortion types