UNIVERSIDADE DA CORUÑA

Learning Evidence of Depression Symptoms via Prompt Induction

Eliseo Bao Anxo Perez David Otero Javier Parapar

IRLab, CITIC, Universidade da Coruña, Spain

Some Context About the Problem...

Mental health services face limited capacity to monitor individuals at risk, yet people increasingly share their experiences on social media. Automatically identifying symptom evidence in this text could support large-scale screening.

Task

Given a sentence s from a Reddit mental health community and a BDI-II symptom c, predict $f : (s, c) \to \{0, 1\}$, whether s contains first-person evidence of c. We address **21 symptoms** from the BDI-II questionnaire using BDI-Sen, a dataset of annotated Reddit sentences.

- Most symptoms have very few positive examples
- Sentences about related symptoms that do not qualify as evidence
- LLMs struggle to apply stable, symptom-specific judgements

What is needed is a way to **distill what counts as evidence for each symptom** into something stable and reusable, without relying on demonstrations at every inference call or on extensive parameter updates.

What is Prompt Induction?

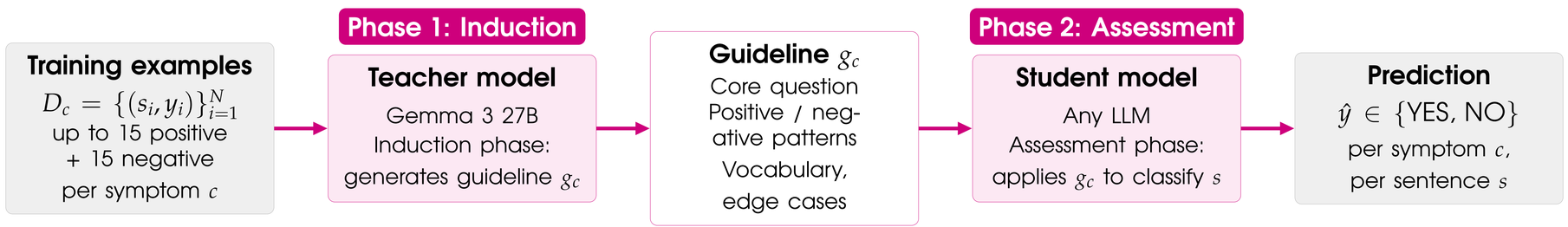
Given a small set of labeled examples, **prompt induction** automatically generates a concise natural-language **guideline** that captures what the task requires. At inference, the model is conditioned on the guideline rather than on the examples themselves.

$$\{(x_i, y_i)\}_{i=1}^{N} \xrightarrow{\text{teacher model}} g \xrightarrow{\text{student model}} \hat{y}$$

The key insight is that a good guideline **compresses** what makes examples positive or negative into an explicit, human-readable specification, stable across inputs and reusable across models.

- **Sample-efficient:** effective from very few labeled examples
- **Interpretable:** human-readable and refineable by domain experts
- **Transferable:** generated once by a teacher, applied by any student
- **Optimization-free:** no parameter updates required at inference

How We Apply It: *Symptom Induction (SI)*



- Guidelines are induced **once** from BDI-Sen training data and reused across all test sentences and student models
- The guideline structure captures: core question, positive and negative patterns, symptom vocabulary, and edge cases
- At inference, the guideline is injected into the system prompt alongside the symptom definition and test sentence

Induction mechanism

The teacher receives up to 15 positive and 15 negative examples per symptom, sampled to include both **hard negatives** (other symptoms) and **soft negatives** (unrelated sentences). It is prompted to produce a guideline that separates relevant from non-relevant evidence, encoding what makes a sentence count as symptom evidence.

Assessment phase

At inference, the induced guideline g_c is injected into the system prompt alongside the BDI-II symptom definition. The model applies the guideline's criteria to each test sentence independently, producing a binary YES/NO prediction. The same guideline is reused for every sentence, giving consistent, reproducible decisions with no demonstrations.

Example: Assessment Prompt for *Sadness*

```
[SYSTEM]
You are a clinical assistant analyzing text for depression symptoms (BDI-II). Determine if the text indicates the person is CURRENTLY experiencing the specified symptom. Answer YES if the text explicitly expresses or clearly implies the symptom now. Answer NO if unrelated, past-only, or a different symptom.

-- Guideline: Sadness --

CORE QUESTION: Does the sentence express feelings of unhappiness, sorrow, or lack of contentment?
SPECIFICALLY IF: contains words like "sad", "unhappy", "down", "sorrow"; expresses worthlessness or self-hate; describes crying or emotional distress; mentions negative outlook or feeling emotionally unfulfilled.
NOT IF: discusses unrelated topics; focuses on fear or nightmares; expresses only mild frustration.
KEY VOCABULARY: sad, unhappy, sorrow, worthless, negative outlook, emotionally unfulfilled, hate myself
TRICKY CASES: distinguish sadness from anxiety (fear-based) or broader depressive experience.
```

Experiments and Results

We compare four strategies on BDI-Sen: zero-shot (**ZS**), in-context learning (**ICL**), supervised fine-tuning (**SFT**), and our proposed **SI**.

Weighted F1 on BDI-Sen. **Bold**: best; underline: second best.

Model	ZS	ICL	SFT	SI
Gemma 3 4B	0.267	0.171	0.242	**0.389**
Gemma 3 12B	0.296	0.268	0.282	**0.367**
Llama 3.2 3B	**0.230**	0.138	0.203	0.229
Llama 3.1 8B	0.227	0.200	0.223	**0.275**
Phi-4 Mini	0.277	0.167	**0.311**	0.277
Phi-4	0.279	0.196	0.305	**0.329**
Qwen3 4B	0.285	0.267	0.299	**0.359**
Qwen3 14B	0.296	0.199	0.263	**0.348**

- SI is best for **6 of 8 models**; Gemma 3 4B + SI (0.389) outperforms all other model/strategy combinations, including larger models with fine-tuning
- **4B model with SI beats 12–14B** models with ZS or SFT
- We also test cross-domain on PsySym, with SI achieving the best raw average F1 and showing that induced guidelines **generalise**

IRLAB INFORMATION RETRIEVAL LAB citic UNIVERSIDADE DA CORUÑA MINISTERIO DE CIENCIA E INNOVACIÓN Financiado por la Unión Europea NextGenerationEU Plan de Recuperación, Transformación y Resiliencia AGENCIA ESTATAL DE INVESTIGACIÓN XUNTA DE GALICIA

SIGIR 2026, The 49th International ACM SIGIR Conference on Research and Development in Information Retrieval. July 20–24, 2026. Melbourne, VIC, Australia.