2ND SMASHING WORKSHOP (19–23 OCT 2026)

Resource-Efficient LLMs for Depression Symptoms Screening: Performance and Limitations in Zero Shot Setting

Scan for the paper

Muhammad Rizwan · Jure Demšar

Sixth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID @ MENTAL.ai), LREC 2026

Co-funded by the European Union

SMASH machine learning for science and humanities postdoctoral program

I FEEL SLOVENIA

UNIVERZA V LJUBLJANI FRI Fakulteta za računalništvo in informatiko

Global burden of depression

322M

people affected worldwide (2023)

5.2%

adult prevalence

women 6.2% · men 4.1%

727K

suicide deaths per year

2021

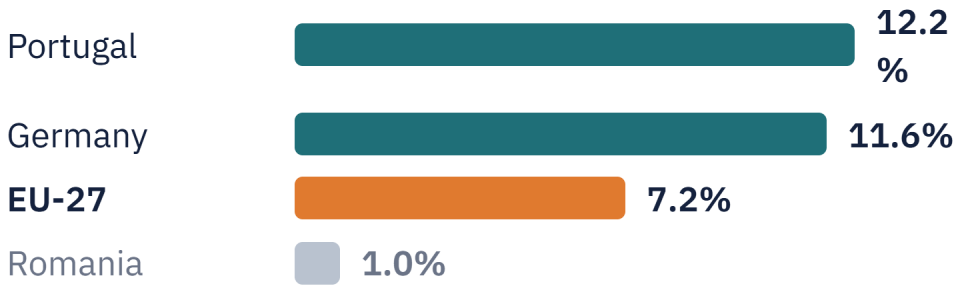
Early detection is a public-health priority.

Source: WHO Depression fact sheet, Sept 2026 (2023 prevalence estimates)

Prevalence and treatment gap

EU PREVALENCE · EHIS 2019

7.2% chronic depression, age ≥ 15



Higher in women in all Member States

MINIMALLY ADEQUATE TREATMENT

1 in 5 high-income countries



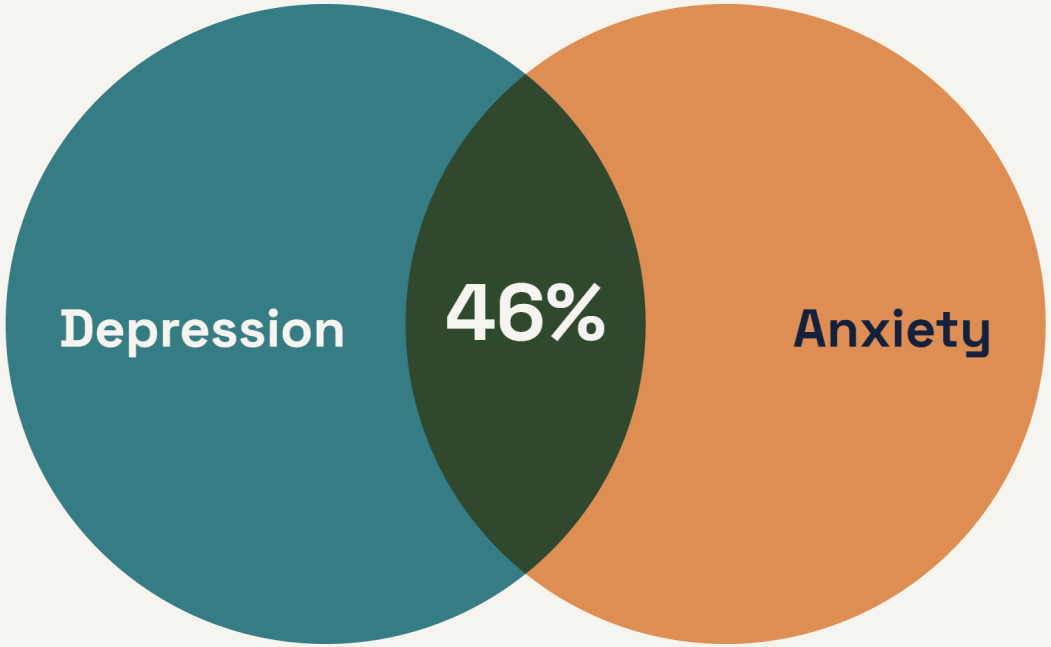
1 in 27 low-income countries



Bottleneck: shortage of clinicians for screening

Sources: Eurostat, European Health Interview Survey 2019 (self-reported) · Thornicroft et al., Br J Psychiatry (2017), WHO World Mental Health Surveys, 21 countries

Psychiatric comorbidity



46% lifetime comorbidity of MDD with anxiety disorders

OTHER FREQUENT COMORBIDITIES

- Substance use disorders
- PTSD
- Chronic somatic illness

⚠ Greater severity · poorer treatment response · overlapping symptomatology

Source: Kessler et al. (2015), Epidemiology and Psychiatric Sciences — WHO World Mental Health Surveys, 74,045 adults, 24 countries (lifetime comorbidity)

Rationale: small open-weight LLMs

Privacy

On-premise inference; data governance

Efficiency

1B–14B parameters; commodity GPUs

Zero-shot

No annotated training data required

Openness

Permissive licences; transparent, auditable weights

RQ Which model maximises clinician agreement — and **which symptoms are systematically missed?**

DSM-5 criteria for MDD

Diagnostic and Statistical Manual of Mental Disorders

Published by

American Psychiatric Association (APA)

DSM-5 · 2013

DSM-5-TR · 2022

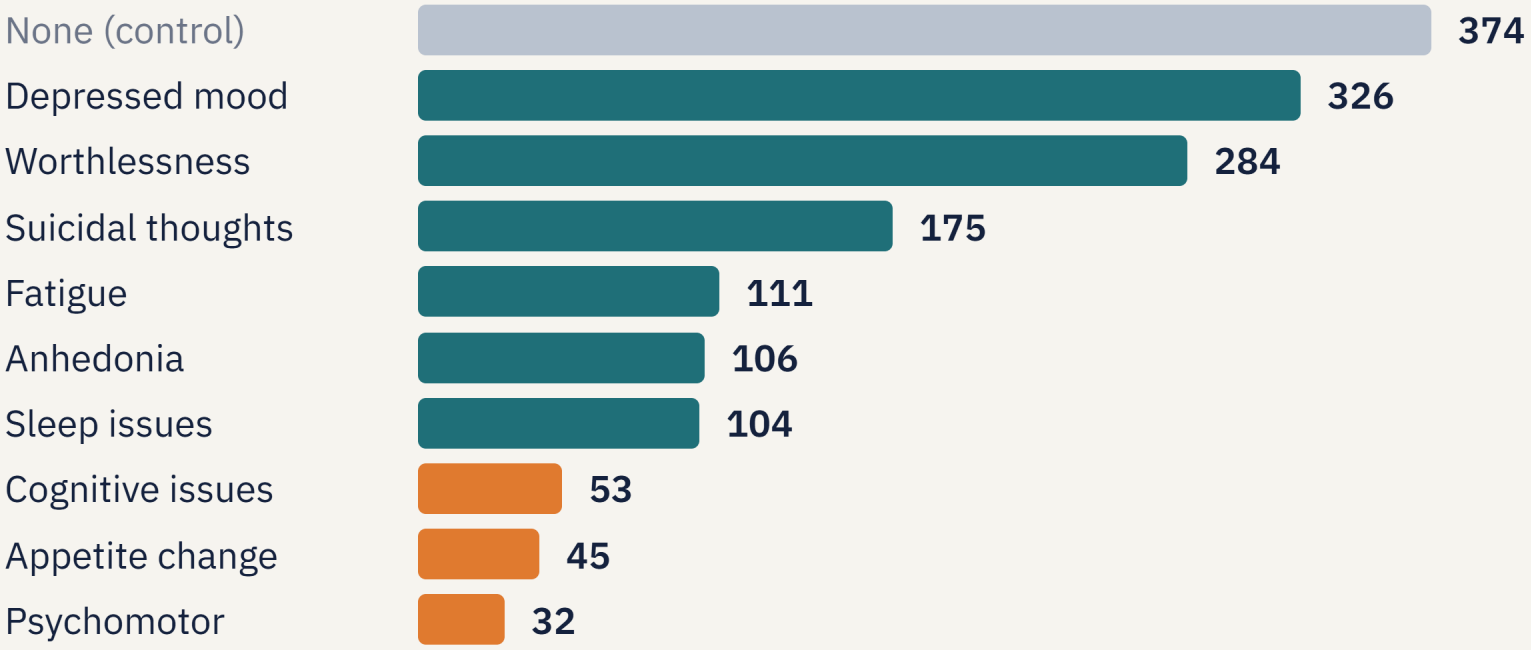
9 DIAGNOSTIC CRITERIA · MAJOR DEPRESSIVE DISORDER

1 Depressed mood	**2 Anhedonia**	3 Appetite / weight change
4 Sleep disturbance	5 Psychomotor change	6 Fatigue / loss of energy
7 Worthlessness / guilt	8 Diminished concentration	9 Suicidal ideation

≥ 5 of 9 criteria **duration ≥ 2 weeks** **≥ 1 core symptom**

Orange = core symptoms · Source: American Psychiatric Association, DSM-5 (2013); DSM-5-TR (2022)

Dataset: ReDSM5 (N = 1,610)



10×

class imbalance (mood vs. psychomotor)

Source: Reddit

sentence-level instances

Expert annotation

licensed clinical psychologist

Single-label subset

129 multi-label instances excluded

ReDSM5 (Bao et al., 2025) · Hugging Face: irlabudc/redsm5 · orange = rare classes (< 60)

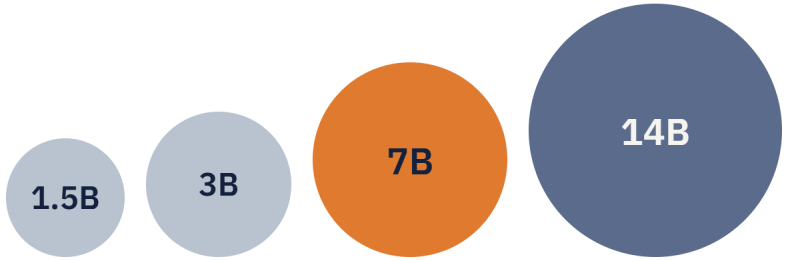
Models: 8 open-weight LLMs

Llama 3.2



Meta

Qwen 2.5



Alibaba · full scaling ladder

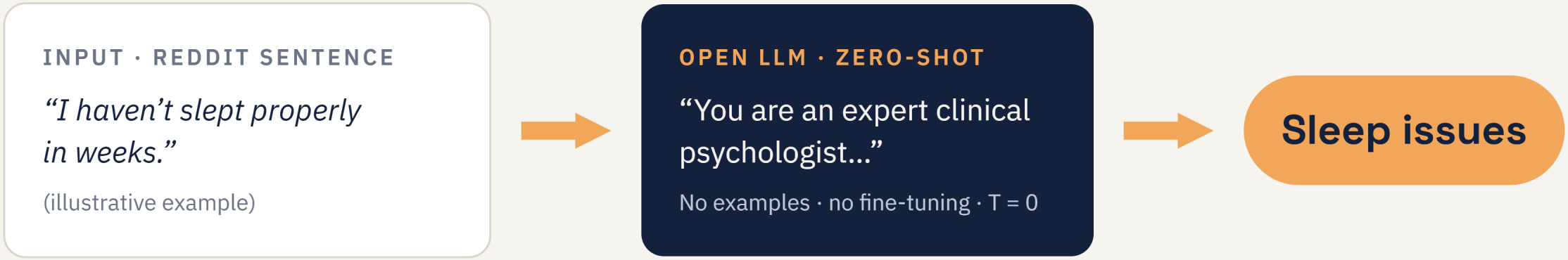
Mistral



7B v0.3 · Nemo 12B

All instruction-tuned · circle size ∝ parameters · orange = best performer

Zero-shot symptom classification



⚠ **Challenge:** single-sentence input — no discourse context, user history or clinical metadata

10-CLASS LABEL SPACE

Depressed mood	Anhedonia	Appetite	**Sleep issues**	Psychomotor
Fatigue	Worthlessness	Cognitive	Suicidal thoughts	None

Dataset: ReDSM5 · 1,610 single-label sentences · annotated by a licensed clinical psychologist

Experimental setup and evaluation

PROMPT

- Persona prompt: expert clinical psychologist (DSM-5)
- ✓ Single-label output: 9 symptoms + NONE
- </> “Output ONLY the category label.”

T = 0 (greedy decoding) | max 15 tokens | 0-shot

Cohen’s κ

Chance-corrected agreement with clinician

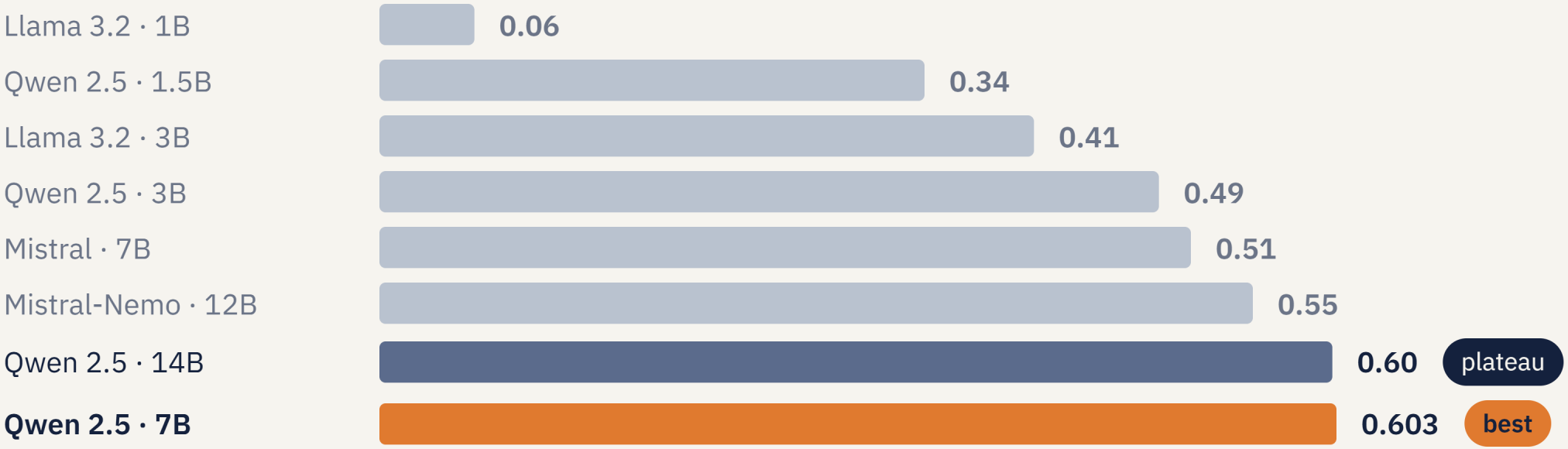
Macro F1

Unweighted mean of per-class F1

Format compliance

Proportion of exact-match label outputs

Agreement plateaus beyond 7B



κ(7B) ≈ κ(14B) → no gain from 2× parameters

Cohen's κ vs. clinician labels · Macro F1: Qwen 7B 0.648, Qwen 14B 0.651

Per-symptom performance: salient vs. core

High F1

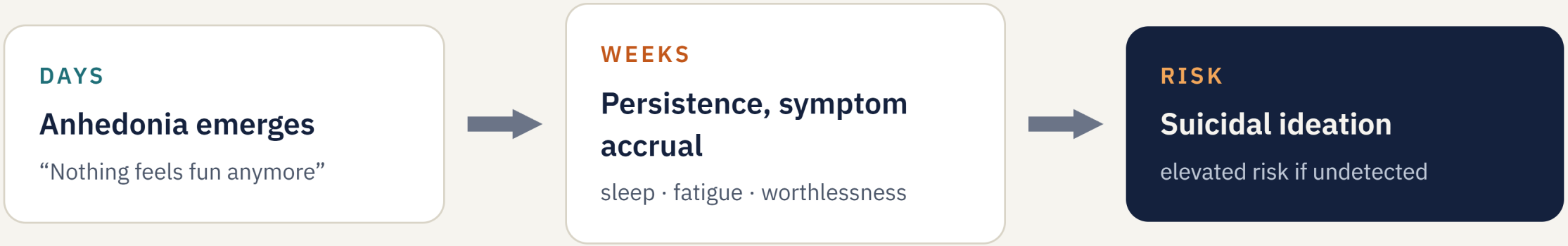
0.89	Suicidal ideation
0.79	Worthlessness
≈0.76	Sleep · Fatigue

⚠ High false-negative rate

88%	Psychomotor
33%	Depressed mood
28%	Anhedonia

Qwen 2.5-7B · per-symptom F1 and false-negative rate

Anhedonia as an early risk marker



Missed in **28%** of cases → **a reduced window for early intervention.**

Source: Ducasse et al. (2018), Depression and Anxiety, 35(5) — meta-analysis · timeline illustrative

Error analysis: dominant confusions

72	Depressed mood	None	Negated or irritable affect → symptom judged absent
66	None	Depressed mood	Lexical cue reliance: transient vs. persistent affect
41	None	Sleep issues	Neutral sleep duration → false positive

⚠ **Positive-prediction bias** → F1(None) = 0.38

Qwen 2.5-7B · true label → predicted label · count of sentences

Qualitative error analysis

FALSE POSITIVE

Missing situational context

Event-bound sadness not distinguished from persistent depressed mood

“I still cry every time I watch this movie.”

FALSE NEGATIVE

Missed hypersomnia

Excessive sleep read as normal behaviour, not a possible depressive symptom

“I slept 15 hours yesterday.”

Core issue: keyword-level matching instead of semantic and contextual understanding

Potential applications

Chatbot safety check

Spot symptoms in messages and point people to help

Clinic triage

Check intake forms first, so urgent cases are seen sooner

Private, local use

Runs on local computers; patient data stays in the clinic

Places with few clinicians

A cheap first screening step

⚠ **Human in the loop:** it helps clinicians; it does not diagnose

Limitations and future work

Single prompt template	Few-shot prompting; prompt ablation
Single-source data (Reddit)	Cross-domain validation
Single-label setting	Multi-label classification
Binary presence only	Severity estimation (mild → severe)
Zero-shot inference only	Parameter-efficient fine-tuning (LoRA)

Thank you

Questions?

Muhammad Rizwan · Jure Demšar

RaPID-6 Workshop @ LREC 2026 (with MENTAL.ai) · pp. 74–81

Read the paper

aclanthology.org/2026.rapid-1.7

Co-funded by the European Union

SMASH machine learning for science and humanities postdoctoral program

I FEEL SLOVENIA

UNIVERZA V LJUBLJANI FRI Fakulteta za računalništvo in informatiko