INSALyon2 at eRisk 2026 Task 1: Safety-First Hybrid Multi-Agent Pipeline for Conversational Depression Detection at eRisk 2026

Notebook for the eRisk Lab at CLEF 2026

Si Hui Ong[1,*], Sumit Sanjay Shinde[1] and Diana Nurbakova[2]

[1]INSA Lyon, 69100 Villeurbanne, France
[2]INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69100 Villeurbanne, France

Abstract

We describe our system for CLEF eRisk 2026 Task 1, *Conversational Depression Detection with LLM Personas*. The task requires a doctor agent to conduct an indirect, natural-language interview with 20 simulated patient personas (Llama-3-8B with per-persona LoRA adapters) and to predict each persona's Beck Depression Inventory–II (BDI-II) total score together with up to four key symptoms. Our approach is a hybrid *agent-to-agent* (A2A) pipeline that orchestrates six specialised components—Stopper, Prober, Template Evidence, Risk Router, Extractor, and Scorer—combining LLM-based question generation and symptom extraction (DeepSeek `deepseek-chat`) with deterministic rule-based stopping, safety routing, and score calibration. Three distinguishing design choices underpin the system: (i) a graded *acute safety ladder* that ensures suicidal-ideation cues are clarified through a five-stage follow-up sequence before the interview terminates; (ii) structured YAML interview banks with group-screen and symptom-drilldown questions that guarantee balanced BDI-II domain coverage; and (iii) configurable run policies that instantiate balanced, recall-oriented, and precision-oriented submission profiles through threshold and stopping-rule variation alone, without changing the underlying model. Internal evaluation on the TalkDep corpus yields a Spearman $\rho = 0.952$ and MAE of 2.46 for our LLM extractor, compared with $\rho = 0.847$ and MAE of 10.36 for a lexical fallback. Post-hoc evaluation against the released ground truth (19 personas submitted) yields ADODL scores of 0.81–0.85; Run 1 achieves the highest average symptom hit rate (ASHR = 0.276) among incomplete submissions in the preliminary results report. Component ablation experiments confirm the contributions of evidence memory, stopping logic, and acute calibration. We release three official runs and discuss the transparency and auditability benefits of the hybrid architecture.

Keywords

depression detection, BDI-II, multi-agent system, conversational AI, LLM

1. Introduction

Depression is the leading cause of disability worldwide, affecting an estimated 280 million people [1]. Despite robust psychometric instruments such as the Beck Depression Inventory–II (BDI-II) [2], screening coverage remains limited by clinician availability, patient reluctance, and geographic barriers. Natural language processing (NLP) offers a scalable alternative: systems can analyse text produced in social media, clinical interviews, or structured dialogues to infer depressive symptomatology [3].

The CLEF eRisk initiative has driven this research agenda since 2017, first through sequential social-media analysis for early risk prediction [4] and more recently through symptom-level BDI-II modelling [5, 6]. The 2025 edition introduced a pilot *Conversational Depression Detection via LLM Personas* track that shifted the paradigm from passive text consumption to active, interactive dialogue: participants deploy doctor agents that interview fine-tuned patient personas and must predict depression severity and key symptoms from the resulting conversation [7]. eRisk 2026 Task 1 extends this pilot into a full shared task with 20 patient personas, weekly LoRA-adapter releases on top of `Meta-Llama-`

CLEF 2026 Working Notes, 21–24 September, Jena, Germany
*Corresponding author.
ong.sihui1@gmail.com (S. H. Ong); sumitsanjayshinde0702@gmail.com (S. S. Shinde); diana.nurbakova@insa-lyon.fr (D. Nurbakova)
0000-0002-6620-7771 (D. Nurbakova)

`3-8B-Instruct` [8], and an explicit evaluation framework targeting BDI-II total scores and symptom identification, detailed thoroughly in both the track working notes [9] and the core LNCS overview [10].

The *agent-to-agent* (A2A) setting introduces challenges that go beyond traditional text classification. First, the doctor agent must *elicit* evidence indirectly—asking "Are you depressed?" is both clinically inappropriate and unlikely to produce nuanced symptom-level signals [11]. Second, the agent must cover all 21 BDI-II dimensions without over-probing personas who show no depressive signs, since unnecessary questioning wastes the limited conversational budget. Third, when a persona expresses suicidal ideation, self-harm intent, or ambiguous fatalistic language (e.g. "I've accepted my fate"), the system must clarify and follow up responsibly before either continuing the interview or classifying the persona. Clinical literature on structured suicide-risk assessment motivates a graded follow-up protocol rather than an immediate stop or a blunt safety question [12].

Recent work on multi-agent mental health systems has explored related design spaces. MAGI [13] coordinates multiple LLM agents to administer psychiatric interviews with structured reasoning. DSM5AgentFlow [14] uses therapist, client, and diagnostician agents to generate and evaluate DSM-5 questionnaire dialogues. AgentMental [15] introduces an interactive multi-agent framework for explainable mental health assessment. EmoScan [16] trains an LLM agent on synthetic clinical interviews to screen for anxiety and depressive disorders with explanations. These systems demonstrate that multi-agent architectures can improve both diagnostic accuracy and transparency, but none address the specific challenge of BDI-II symptom-level inference from an open-ended conversational interview with persona-specific LoRA-adapted patient simulators.

Several systems from the eRisk 2025 pilot adopted prompt-engineering strategies, using diverse LLMs to generate BDI-II-aligned questions and structured JSON outputs [17, 18]. While effective as baselines, these approaches typically rely on a single LLM for all functions—question generation, risk assessment, symptom scoring, and stopping decisions—making it difficult to audit or override individual components.

We take a different approach. Our system separates *language understanding* (where LLMs excel) from *safety progression, interview completeness, and score calibration* (where explicit rules provide transparency and auditability). The result is a **hybrid multi-agent pipeline** with six components, each independently testable and interpretable:

1. **Risk Router**: classifies the current interview cluster (e.g. `AcuteSafety`, `CoreDepression`) and selects the next probing strategy.
2. **Prober**: generates the next doctor question via DeepSeek [19], guided by topic hierarchy, YAML interview banks, and a red-flag follow-up policy.
3. **Template Evidence**: scores each patient turn against symptom templates and a risk lexicon using either sentence-transformer embeddings [20, 21] or token-overlap matching.
4. **Extractor**: a second DeepSeek call that maps dialogue history to 21 BDI-II symptom scores (0–3), with explicit instructions for suicide language, fatalistic framing, and evidence-count gating.
5. **Stopper**: a rule-based module that balances group-screen coverage, acute-ladder progress, early positive-framing exits, and minimum-depth requirements.
6. **Scorer**: sums symptom scores with acute-risk calibration that can raise totals when suicidal-ideation signals, ladder coverage, and risk-buffer evidence align.

We submit three runs that differ only in *threshold and stopping parameters*—balanced, recall-oriented, and precision-oriented—demonstrating that a single architecture can cover multiple operating points without retraining or prompt modification.

The rest of this paper is organised as follows. Section 2 summarises the task definition and constraints. Sections 3–4 detail the system architecture, with emphasis on the safety ladder and interview bank design. Section 5 describes the three run policies. Section 6 presents internal evaluation on the TalkDep corpus [22], component ablations, and symptom difficulty analysis. Section 7 discusses the design rationale and positions the system relative to prior work. Section 8 addresses limitations and ethical considerations. Section 9 concludes.

Table 1
System components and their roles in the hybrid A2A pipeline.

Component	Role	Type	Key mechanism
Risk Router	Classify interview cluster	LLM + fallback rules	Six clusters; priority: `AcuteSafety` > others
Prober	Generate next doctor question	LLM (DeepSeek)	Topic hierarchy + YAML banks + red-flag policy
Template Evidence	Turn-level risk features	Embedding / lexical	Symptom templates + risk lexicon matching
Extractor	Map dialogue to BDI-II scores	LLM (DeepSeek)	21-symptom JSON; post-parse safety overrides
Stopper	Continue vs. CLASSIFY decision	Rule-based	Group coverage + acute ladder + positive framing
Scorer	Final BDI-II total + key symptoms	Rule-based	Sum + acute calibration; top-4 symptom selection

2. Task Definition

eRisk 2026 Task 1 (Conversational Depression Detection with LLM Personas) requires participants to interact with 20 simulated patient personas and predict, for each persona, a BDI-II total score (0–63) and up to four key depressive symptoms.

Patient personas. Each persona is a per-persona LoRA adapter [23] applied to `Meta-Llama-3-8B-Instruct` [8], released incrementally on Hugging Face. Our official submissions cover personas 1–19 ($n = 19$); persona 20 was not submitted. Personas span the full BDI-II severity spectrum and respond with a frozen system prompt.

Submission format. For each of up to three runs, teams submit two JSON files per persona: `interactions_run{N}.json` (full conversation logs) and `results_run{N}.json` (predicted BDI-II score and key symptoms). At most one run may be manual (human-assisted).

Evaluation. The organisers evaluate (i) accurate identification of depressive symptoms present in the persona and (ii) the overall depression level following BDI-II standards [9, 10]. After the submission deadline, reference BDI-II scores and key symptoms for all 20 personas were released to participants for post-hoc analysis. Each persona index maps directly to the evaluation files provided by the organizers.

3. System Architecture

The **orchestrator** runs a turn loop: at each step it invokes the Stopper to decide whether to continue or classify, then either the Prober (to generate the next question) or the Scorer (to produce the final prediction). After each patient reply, the Template Evidence module and the Extractor update their respective signal stores.

3.1. Risk Router and Cluster Routing

The Risk Router assigns the current interview state to one of six **clusters**: `AcuteSafety`, `HopelessWorthless`, `CoreDepression`, `VegetativeCognitive`, `BehavioralArousal`, and `GeneralCheckin`. When the DeepSeek API is available, classification uses a constrained LLM call; otherwise, the system falls back to **lexical pattern matching** over recent patient utterances.

Priority ordering. The classifier checks clusters in strict priority: if the patient indicates *imminent self-harm, intent, plan, or immediacy language*, the cluster is forced to `AcuteSafety`. If resigned or fatalistic language dominates without clear imminence, `HopelessWorthless` is selected. Otherwise, the best non-risk cluster is chosen based on keyword and context matching.

Each cluster maintains a **question bank** with at least five prepared probes. Examples include: “You said this might happen soon—what has made things feel this immediate right now?” (`AcuteSafety`);

"What used to lift your mood before, and does any of that still help now?" (`CoreDepression`); "How has your sleep been affecting your day-to-day energy lately?" (`VegetativeCognitive`).

Acute euphemism patterns. A curated list of euphemistic or ambiguous phrases—e.g. "end soon", "leave earth", "won't be here", "waiting for the timer", "no one would notice if I wasn't here"—triggers `AcuteSafety` routing before general lexical rules are evaluated. Ambiguous cues (e.g. "disappear", "just existing") require two or more hits to contribute to the acute signal.

3.2. Acute Safety Ladder

When acute safety cues are detected, the system activates a graded **five-stage follow-up ladder** rather than immediately terminating or bluntly asking about suicide. This five-stage follow-up structure mirrors the clinical hierarchy derived from validated screening tools, notably the Patient Health Questionnaire-9 (PHQ-9) Item 9 protocol [12], and standard emergency medicine triage pathways. The ladder sequentially quantifies severity from cognitive ideation down to physical access parameters, gathering progressively more specific information while maintaining conversational rapport.

1. **Intent clarification.** "When you say this will end soon, are you talking about ending your life?"
2. **Plan concreteness.** "Have you thought about specific steps or a specific way you would do it?"
3. **Timeline imminence.** "Have you thought about when you might do this?"
4. **Means access.** "Do you currently have access to what you would use?"
5. **Protective factors.** "What has stopped you so far from acting on these thoughts?"

Each stage has **substring markers** (e.g. "ending your life", "specific steps") that the system uses to detect whether a stage has already been covered in the conversation. The Stopper tracks `acute_ladder_progress` and will not allow the interview to terminate until the policy-mandated minimum number of ladder steps has been reached (3 for run1, 2 for run2, 4 for run3; see Section 5).

3.3. Prober: Indirect Question Generation

The Prober generates each doctor question via a DeepSeek `deepseek-chat` call with a fixed system prompt that enforces key constraints:

- Output exactly one natural question per turn.
- Never use the words "depression" or "mental health" or ask "Are you depressed?".
- Move **general → specific** within each functional domain.
- **If the patient says something concerning or ambiguous** (e.g. "accepted my fate", "know how this ends", "made my peace"), ask a **follow-up before changing topic**.

Question generation is further guided by three mechanisms:

Topic hierarchy. Six topics—General Mood, Physical, Motivation, Self Outlook, Cognitive, and Behavioural/Emotional—define keyword sets, mapped BDI-II symptoms, opening questions, and follow-up questions. The orchestrator tracks which topics have been visited and preferentially selects under-covered ones.

YAML interview banks. Two manually curated question banks provide structured interview content. These banks were designed prior to runtime by clinical mapping of the official BDI-II diagnostic criteria to natural language indicators, ensuring human alignment and preventing automatic generation drift:

- *Group screen questions* cover four functional groups (Affective, Executive, Somatic, Cognitive), each with indirect screening questions targeting specific BDI-II symptoms. For example, "How have you been feeling lately, overall?" targets Sadness and Loss of Pleasure; "Is it harder to focus or concentrate than it used to be?" targets Concentration Difficulty.
- *Symptom drilldown questions* narrow the assessment per symptom after screening. For example, for Sadness: "When you say things have been rough, is that most days or just now and then?"; for Suicidal Thoughts: "When things feel really heavy, what thoughts show up about getting through it?"

The policy enforces a minimum of `min_questions_per_group_screen` screen questions per group (default 3) before early stopping is allowed, and caps drilldown questions per symptom at `max_questions_per_symptom` (default 3).

Bank follow-up. After each bank question, an additional LLM call generates a brief contextualised follow-up grounded in the patient's last answer (e.g. "What part of that has been the hardest for you day to day?"), maintaining conversational flow.

3.4. Template Evidence and Risk Buffer

Each patient turn is scored against two knowledge resources:

Symptom templates. Short first-person paraphrases (e.g. "I feel sad", "I am discouraged about my future", "I have thoughts of killing myself") are used for embedding similarity (via `all-MiniLM-L6-v2` [20, 21]) or token-overlap matching when embeddings are disabled. The highest-scoring template determines the turn's primary symptom signal.

Risk lexicon. Two phrase lists—`acute_safety` ("kill myself", "want to die", "leave earth", ...) and `hopeless_worthless` ("nothing matters", "pointless", "better without me", ...)—boost the turn risk score.

Composite scoring. `compute_turn_risk_score` starts from the maximum template match score, adds +0.5 for any acute lexicon hit, +0.2 for a hopeless hit, and +0.2 if the matched template symptom is Suicidal Thoughts or Wishes, then clips to 1.0.

A **risk buffer** of configurable size (default 6) retains the highest-risk turns with recency weighting (weight 0.15) for downstream use in stopping logic and acute calibration.

3.5. Evidence Memory

The system maintains an optional **evidence memory** that stores patient utterances as sentence-transformer embeddings. When generating the next question, the Prober can retrieve the top-k most relevant prior patient statements (cosine similarity > 0.2) to ground its probe. If the embedder is unavailable, a token-overlap ranking is used instead. This mechanism prevents the Prober from asking about topics the patient has already addressed and supports more targeted follow-up.

4. Symptom Extraction and Scoring

4.1. LLM-Based Extractor

The Extractor uses a dedicated DeepSeek call that outputs a JSON object with the exact 21 BDI-II symptom keys and scores $\in \{0, 1, 2, 3\}$. The system prompt contains detailed instructions for handling:

- **Suicidal language:** explicit plans or intent language must receive high scores even if expressed calmly.
- **Fatalistic framing:** phrases like "accepted my fate" or "I know how this ends" with paradoxically positive tone must not be under-scored.
- **Genuine positive health:** truly positive somatic reports should not inflate sleep or fatigue scores.

Post-parse rules add a rule-based correction layer:

- *Hopeful markers* ("I hope things get better", "I'm trying") without acute phrases cap Pessimism, Worthlessness, and Suicidal Thoughts at 1 if the model scored higher.
- *Evidence-count gating*: for symptoms including Agitation, Indecisiveness, Worthlessness, and Concentration Difficulty, if the score exceeds 1 but fewer than two patient messages match domain patterns, the score is clamped to 1.
- *Safety overrides*: phrases such as "gonna end it", "sleep forever", or "happy to die" force Suicidal Thoughts ≥ 3.

Table 2
Key parameters distinguishing the three submission runs. All other parameters (banks, follow-up, embeddings) are identical.

Parameter	Run 1 *balanced*	Run 2 *recall*	Run 3 *precision*
`severe_threshold`	24	20	28
`control_threshold`	5	6	4
`min_exchanges`	10	9	11
`acute_ladder_steps`	3	2	4
`acute_boost_floor`	50	46	54
`consec. confirm.`	2	1	3

Fallback extractor. When no API is available, a keyword-based fallback maps fixed phrase lists to symptom indices (e.g. mood words → Sadness; self-harm language → Suicidal Thoughts). This path is not the primary production mechanism but serves as a robustness guarantee.

4.2. Stopper: Stopping Logic

The Stopper enforces several conditions before allowing classification:

1. **Minimum exchanges:** at least `min_exchanges_before_stop` turns (10 for run1, 9 for run2, 11 for run3).
2. **Group screen completeness:** each of the four functional groups must have received at least `min_questions_per_group_screen` screening questions.
3. **Minimum distinct group coverage:** at least `min_groups_before_stop` (default 3) groups must have been probed.
4. **Acute risk gating:** if acute cues are present, stopping is delayed until `acute_ladder_progress` $\geq$ `required_acute_ladder_steps`.
5. **Core domain coverage:** extractor signal or interview coverage must span sleep/energy, interest/pleasure, self-view, and future outlook before early exit is permitted.
6. **Positive framing early stop:** if the patient's first few turns contain globally positive phrases ("doing well", "feeling good", "I'm good", ...), stopping can trigger sooner with a relaxed threshold (`positive_framing_threshold`, default 8) to avoid over-probing non-depressed-like personas.

4.3. Scorer: BDI-II Total and Key Symptoms

The Scorer sums symptom scores (clamped to $[0, 3]$ per item) to produce a raw total $\in [0, 63]$. An **acute-risk calibration** step can raise the total when three conditions align:

1. The `has_acute_signal` flag is true.
2. The Suicidal Thoughts symptom signal is high.
3. The `acute_ladder_progress` meets the policy's `required_acute_ladder_steps` threshold.

The calibration applies policy-defined floors (`acute_boost_floor`, `moderate_acute_boost_floor`, `mild_acute_boost_floor`) and the risk buffer's maximum template risk.

Key symptom selection. Up to four symptoms are selected by choosing the highest non-zero scores, with ties broken in canonical BDI-II order. The selection is validated via `validate_key_symptoms` which ensures symptom names match the official 21-item list.

5. Run Policies

We submit three runs that differ only in *stopping and threshold parameters*, not in model or prompt design. Table 2 summarises the key differences.

Table 3
Extractor performance on TalkDep (11 personas).

Metric	LLM Extractor	Lexical Fallback
MAE	2.455	10.364
RMSE	3.705	13.403
Spearman ρ	0.952	0.847

Table 4
Per-persona BDI-II scores for selected TalkDep personas (reference, LLM extractor, lexical fallback).

Persona	Ref.	LLM	Fallback
Maria	40	40	15
Marco	38	41	16
Linda	28	29	10
Laura	23	33	7
James	22	22	10
Gabriel	13	12	4
Noah	5	7	5

Run 1 (balanced) uses moderate thresholds and represents the default operating point. **Run 2 (high recall)** lowers the severe threshold, requires fewer acute ladder steps, and allows a more permissive control threshold—tilting toward not missing depression or safety signals at the cost of potential false positives. **Run 3 (high precision)** raises thresholds, requires more ladder steps and longer minimum interviews, and uses higher acute calibration floors—tilting toward fewer false positives at the cost of potentially missing borderline cases.

This design demonstrates that *a single architecture can span a meaningful precision–recall operating range* through policy variation alone.

6. Internal Evaluation

Following the organisers' release of official reference labels [9, 10], we report both development proxies (TalkDep, component ablation) and post-hoc evaluation on our three submitted runs against the released ground truth.

Important caveat. TalkDep and ablation numbers remain *development proxies*. Post-hoc scores in Section 6.4 reflect our submitted predictions and should be interpreted alongside the official shared-task evaluation protocol. The component ablation in particular reflects an exploratory single-persona mock configuration and is presented for *methodology illustration* rather than definitive quantitative claims.

6.1. TalkDep Corpus Calibration

The TalkDep corpus [22] provides 11 clinically grounded simulated personas with reference BDI-II scores spanning minimal to severe depression (range: 5–40). We ran our Extractor on final conversation transcripts and compared against reference scores.

Table 3 shows that the LLM-based extractor dramatically outperforms the lexical fallback across all metrics: a fourfold reduction in MAE and a much stronger rank correlation. The fallback's high RMSE is driven by systematic under-scoring of cognitive and interpersonal symptoms (e.g. Self-Criticalness, Loss of Pleasure, Loss of Interest) that lack explicit keyword anchors.

Per-persona examples. Table 4 shows selected personas spanning the severity range. The LLM extractor closely tracks reference scores (Maria: ref 40, pred 40; James: ref 22, pred 22), while the fallback consistently under-estimates severe personas (Marco: ref 38, fallback 16) and occasionally over-estimates mild ones.

Table 5
Component ablation results (illustrative, single-persona mock run). MAE is computed against the baseline score.

Variant	Score	Turns	$\lvert\Delta\rvert$
Baseline	16	10	—
No memory	23	20	7
No template risk	16	20	0
No stopper	12	20	4
Fallback extractor	2	20	14
No acute calibration	12	20	4

6.2. Component Ablation

To understand each component's contribution, we ran a controlled ablation study with six variants:

1. **Baseline**: full system.
2. **No memory**: evidence memory disabled; Prober has no access to prior patient statements.
3. **No template risk**: template matching and risk lexicon scoring disabled.
4. **No stopper**: stopping logic removed; interview runs to maximum turn limit.
5. **Fallback extractor**: LLM extractor replaced with keyword-based fallback.
6. **No acute calibration**: acute-risk score calibration disabled.

Table 5 reports illustrative results.
Key observations:

- *Fallback extractor* produces the largest score deviation ($|\Delta| = 14$), confirming the LLM extractor's critical role.
- *No memory* raises the score by 7 points and doubles the turn count, indicating that without evidence memory the Prober fails to recognise already-covered topics and over-probes, which inflates the extractor's cumulative signal.
- *No stopper* and *No acute calibration* both reduce the score by 4 points: without stopping logic the system runs to maximum depth, losing the coverage-based stopping signal; without calibration, acute personas are under-scored.
- *No template risk* has no effect on the final score in this mock configuration but doubles the turn count, showing that template evidence influences stopping (through risk buffer) more than extraction.

These results, though illustrative, confirm that the hybrid design distributes critical functions across components rather than concentrating them in a single LLM call.

6.3. Symptom Difficulty Analysis

We analysed extractor disagreement between the LLM and fallback extractors across TalkDep personas to identify the "hardest" symptoms—those where the two methods disagree most frequently.

Table 6 reveals an instructive pattern. For *Self-Criticalness* and *Loss of Pleasure*, the LLM extractor detects signal in over 80% of personas while the fallback finds almost none—these symptoms are expressed through nuanced language that lacks explicit keyword anchors. Conversely, *Sadness* shows higher fallback prevalence (0.73 vs. 0.36), suggesting the LLM extractor may under-score explicit sadness language that the keyword method captures. This complementarity motivates the hybrid design: lexical features feed the template evidence and risk buffer (contributing to stopping and calibration), while the LLM extractor handles nuanced inference.

Table 6
Top-5 symptoms by extractor disagreement rate on TalkDep ($n = 11$ personas). Prevalence is the fraction of personas where each extractor scores > 0.

Symptom	Disagr. rate	LLM / Fallback avg. score
Self-Criticalness	0.91	1.82 / 0.09
Loss of Pleasure	0.82	1.64 / 0.00
Sadness	0.82	1.00 / 1.45
Tiredness or Fatigue	0.82	2.00 / 1.82
Changes in Sleeping	0.82	1.18 / 0.91

Table 7
Official Task 1 metrics reproduced on INSALyon-2 submissions ($n = 19$ personas; persona 20 not submitted). Organiser preliminary values in parentheses where available.

Run	ADODL	DCHR	ASHR
Run 1 (balanced)	0.8145 (0.8145)	0.2105 (0.2105)	0.2763 (0.2763)
Run 2 (recall)	0.8463 (0.8463)	0.3158 (0.3158)	0.2368 (0.2500)
Run 3 (precision)	0.8496 (0.8496)	0.3158 (0.3158)	0.2105 (0.2237)

6.4. Official Ground-Truth Evaluation

The organisers released reference BDI-II totals and up to four key symptoms per persona after the submission window closed (May 2026) and published preliminary Task 1 results for team **INSALyon-2** (19 personas per run). We reproduce their metrics on our submitted `outputs/submission/` files by matching each persona $p \in \{1, \ldots, 19\}$ directly to ground-truth labels, and normalising predicted symptom strings with the official synonym map.

Following the preliminary report, we compute: **ADODL** (Average Difference between Overall Depression Levels), $\text{ADODL} = \frac{1}{|U|}\sum_u (63 - |\text{ADL}_u - \text{EDL}_u|)/63$; **DCHR** (Depression Category Hit Rate), the fraction of personas whose predicted BDI-II total falls in the same severity band as the reference (minimal 0–9, mild 10–18, moderate 19–29, severe 30–63); and **ASHR** (Average Symptom Hit Rate), the mean fraction of the four reference key symptoms correctly identified after normalisation.

Table 7 shows that our reproduced scores match the organisers' preliminary report exactly for ADODL and DCHR; ASHR differs by at most 0.013 on runs 2–3 due to minor symptom-string normalisation details. Run 1 attains the *highest ASHR among all 19-persona submissions* in the preliminary incomplete-submission table (0.2763). Run 3 achieves our best ADODL (0.8496).

Although raw BDI-II MAE remains moderate (9.5–11.7 points on this subset), ADODL is high because errors are normalised against the 0–63 scale: a 10-point miss still yields $(63 - 10)/63 \approx 0.84$. Our system therefore tracks overall depression level well, with Run 1 particularly strong on symptom identification (ASHR), consistent with the organisers' feedback that our results are accurate relative to other partial submissions. Latency-aware scores (LDCHR/LASHR) are lower in the preliminary report because our runs average 23.4 doctor messages per conversation—a deliberate trade-off for structured group screening and bank follow-ups.

6.5. Probe-Cap Behaviour

We analysed submission interactions to verify that the system respects configured probing caps. The probe-cap analysis counts how often each symptom, group, and topic was targeted versus policy maxima. Across runs, we found zero cap violations in the final submissions, confirming that the interview bank and group-switch-on-saturation logic correctly bounds per-symptom and per-group probing.

7. Discussion

7.1. Why Hybrid Over End-to-End?

A monolithic LLM-based system (single model for questioning, extraction, stopping, and scoring) is simpler to implement but harder to audit, debug, and control. Our experience building and testing the system revealed several failure modes that motivated the hybrid design:

- **LLM optimism bias:** general-purpose LLMs tend to under-score suicidal ideation when the patient expresses it calmly or uses euphemisms. Our extractor's post-parse safety overrides and the explicit instructions for paradoxical framing directly counter this.
- **Coverage imbalance:** without structured interview banks, the Prober would repeatedly ask about mood and sleep—the most salient topics—while neglecting cognitive and self-evaluation domains. The group-screen completeness requirement forces balanced BDI-II coverage.
- **Over-probing controls:** a purely LLM-driven stopper struggles to know when "enough evidence" has been gathered. Our rule-based stopper with positive-framing early exit efficiently curtails interviews for non-depressed personas.

7.2. Safety as a First-Class Concern

The acute safety ladder is, to our knowledge, the first structured suicide-risk follow-up protocol integrated into a conversational depression detection system for a shared task. While the system operates on simulated personas (not real patients), we argue that building safety-aware interviewing into the architecture from the start is essential for responsible development of such technologies. The ladder ensures that ambiguous or euphemistic self-harm language is always clarified before the system either continues probing or produces a final classification.

7.3. Positioning Relative to Prior Work

Compared to prompt-engineering approaches from eRisk 2025 [17, 18], our system adds explicit structure (YAML banks, topic hierarchy) and safety routing. Compared to multi-agent frameworks like MAGI [13] and DSM5AgentFlow [14], our system targets the specific A2A setting of eRisk (per-persona LoRA patients, BDI-II scoring) and introduces the acute ladder and run-policy variation. Compared to recent work on DeepSeek for depression detection [24], our contribution is the integration of DeepSeek into a multi-component pipeline with rule-based guardrails rather than using it as a standalone classifier.

7.4. Transparency and Auditability

Every decision in the pipeline is traceable: the Risk Router logs its cluster classification, the Prober logs which bank or topic it drew from, the Template Evidence logs per-turn risk scores, the Extractor logs raw and post-processed symptom scores, the Stopper logs which conditions triggered or blocked stopping, and the Scorer logs the calibration pathway. This audit trail enables fine-grained error analysis and targeted improvements—a practical advantage over opaque single-model systems.

8. Limitations and Ethical Considerations

Label access. Official reference labels were released after the submission deadline [9, 10]; Section 6.4 reports post-hoc scores on our available submissions. TalkDep metrics (Section 6.1) remain development proxies and may not fully predict performance on the eRisk personas.

API variability. DeepSeek outputs can drift with API updates; exact scores may differ across dates unless conversation logs are frozen.

Ablation scope. The checked-in component ablation reflects a quick mock, single-persona configuration. It illustrates the methodology but should not be treated as definitive quantitative evidence. We plan to rerun ablations under a locked multi-persona protocol for the final camera-ready version.

Embedding dependency. Template and evidence-memory paths depend on local embedding availability. The system degrades gracefully to token-overlap matching, but the quality of risk scoring and memory retrieval may be reduced.

Not a clinical device. This system is a research prototype designed for simulated patient personas within a shared-task evaluation. It has not been validated for use with real patients and should not be deployed in clinical settings without extensive further validation, ethical review, and regulatory approval [25].

Safety limitations. While the acute ladder adds structure to safety-relevant conversations, it cannot guarantee appropriate handling of all possible crisis expressions. The system's responses are generated by an LLM and may occasionally produce inappropriate or insufficient safety responses.

9. Conclusion

We presented a hybrid multi-agent pipeline for conversational depression detection in eRisk 2026 Task 1. The system combines LLM-based question generation and symptom extraction (DeepSeek) with rule-based stopping, safety routing, and score calibration to produce transparent, auditable BDI-II predictions. Three key contributions distinguish our approach:

1. A **graded acute safety ladder** for structured suicide-risk clarification in conversational interviewing.
2. **Structured YAML interview banks** with group-screen and symptom-drilldown questions ensuring balanced BDI-II domain coverage.
3. **Configurable run policies** that span a meaningful precision–recall range through threshold variation alone.

Internal evaluation on TalkDep demonstrates strong rank correlation ($\rho = 0.952$) and low absolute error (MAE = 2.46) for the LLM extractor. Component ablations confirm that each subsystem contributes to the final prediction quality. Symptom difficulty analysis reveals complementary strengths of LLM and lexical extraction, motivating the hybrid design.

We submit three runs (balanced, recall-oriented, precision-oriented). Post-hoc evaluation against the released ground-truth labels (Section 6.4) complements our TalkDep development analysis; our reproduced ADODL scores (0.81–0.85 on 19 personas) align with the organisers' preliminary INSALyon-2 report.

Acknowledgments.

We thank the eRisk 2026 organisers—Javier Parapar, Anxo Perez, Xi Wang, and Fabio Crestani—for designing and maintaining the shared task, releasing the TalkDep patient simulation framework, and hosting the evaluation infrastructure. We also acknowledge the DeepSeek team for providing API access used in our Prober and Extractor components.

A. BDI-II Symptom Mapping

Table 8 lists the 21 BDI-II symptoms in canonical order together with their assigned functional group.

B. Run Policy Configuration

Table 9 shows the complete set of policy parameters for each run.

Table 8
The 21 BDI-II symptoms and their functional group assignments used in the system.

#	Symptom	Group
1	Sadness	Affective
2	Pessimism	Affective
3	Past Failure	Executive
4	Loss of Pleasure	Cognitive
5	Guilty Feelings	Executive
6	Punishment Feelings	Executive
7	Self-Dislike	Affective
8	Self-Criticalness	Executive
9	Suicidal Thoughts or Wishes	Affective
10	Crying	Somatic
11	Agitation	Somatic
12	Loss of Interest	Cognitive
13	Indecisiveness	Executive
14	Worthlessness	Affective
15	Loss of Energy	Somatic
16	Changes in Sleeping Pattern	Somatic
17	Irritability	Somatic
18	Changes in Appetite	Somatic
19	Concentration Difficulty	Cognitive
20	Tiredness or Fatigue	Somatic
21	Loss of Interest in Sex	Cognitive

Table 9
Full run policy configuration from `run_policies.yaml`.

Parameter	Run 1 (balanced)	Run 2 (recall)	Run 3 (precision)
`prober_temperature`	0.25	0.25	0.25
`control_threshold`	5	6	4
`severe_threshold`	24	20	28
`consecutive_confirmations`	2	1	3
`required_acute_ladder_steps`	3	2	4
`acute_boost_floor`	50	46	54
`moderate_acute_boost_floor`	44	40	48
`mild_acute_boost_floor`	38	34	42
`min_exchanges_before_stop`	10	9	11
`max_questions_per_symptom`	3	3	3
`max_questions_per_group`	6	6	6
`min_groups_before_stop`	3	3	3
`min_questions_per_group_screen`	3	3	3
`max_drilldown_questions_total`	24	24	24
`bank_followup_enabled`	true	true	true
`group_screen_enabled`	true	true	true
`symptom_drilldown_enabled`	true	true	true
`group_switch_on_saturation`	true	true	true

C. Example Conversation Excerpt

The following is a representative excerpt from a development run with a mock persona exhibiting depressive signals. Turn labels: **D** = Doctor (our system), **P** = Patient (persona).

D: How have you been feeling lately?
P: Uh… not great, honestly.

D: Can you tell me a bit more about what “not great” feels like?
P: Uh… not great, honestly.
D: How’s your sleep been lately?
P: Uh… not great, honestly.
D: What kinds of things have you been doing with your time?
P: I don’t know. Things are just… hard.
D: You said things are hard—has that made it tough to get started on things you usually enjoy?
P: Nothing really interests me anymore.
D: When you say nothing interests you anymore, how does that make you feel about yourself?
P: Uh… not great, honestly.
D: When you say things feel “not great,” does that colour how you see the future?
P: Uh… not great, honestly.

Note: This mock persona produces limited response variety; real LoRA-adapted personas exhibit richer conversational behaviour.

References

[1] World Health Organization, Depressive disorder (depression): Fact sheet (2023). https://www.who.int/news-room/fact-sheets/detail/depression.

[2] A. T. Beck, R. A. Steer, G. K. Brown, Manual for the Beck Depression Inventory–II, Psychological Corporation, San Antonio, TX, 1996.

[3] M. A. Kuhail, et al., Screening for depression using natural language processing: Literature review, Interactive Journal of Medical Research 13 (2024) e55067. doi:10.2196/55067.

[4] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2016), volume 9822 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 28–39. doi:10.1007/978-3-319-44564-9_3.

[5] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet, in: Proceedings of the Working Notes of CLEF 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[6] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet, in: Proceedings of the Working Notes of CLEF 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[7] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: Proceedings of the Working Notes of CLEF 2025, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[8] A. Dubey, A. Jauhri, A. Pandey, et al., The Llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[9] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and adhd (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, 21–24 September, 2026, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2026.

[10] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and adhd, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 17th International Conference of the CLEF Association, CLEF 2026, Jena, Germany, September 21–24, 2026, Proceedings, volume To be published of *Lecture Notes in Computer Science*, Springer, 2026.

[11] J. C. Tolentino, S. L. Schmidt, DSM-5 criteria and depression severity: Implications for clinical practice, Frontiers in Psychiatry 9 (2018) 450. doi:10.3389/fpsyt.2018.00450.

[12] K. Kroenke, R. L. Spitzer, J. B. W. Williams, The PHQ-9: Validity of a brief depression severity

measure, Journal of General Internal Medicine 16 (2001) 606–613. doi:10.1046/j.1525-1497.2001.016009606.x.

[13] G. Bi, Z. Chen, Z. Liu, X. Xiao, Y. Xie, Y. Zhao, L. He, M. Huang, MAGI: Multi-agent guided interview for psychiatric assessment, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025.

[14] C. Erden, B. Dur, Ö. Uzuner, Multi-agent LLM workflow for counseling and explainable mental health diagnosis, arXiv preprint arXiv:2508.11398 (2025).

[15] G. Bi, et al., AgentMental: An interactive multi-agent framework for explainable mental health assessment, arXiv preprint arXiv:2508.11567 (2025).

[16] Y. Kim, S. Park, S. Kim, et al., Enhanced large language models for effective screening of emotional disorders, Communications Medicine 5 (2025) 136. doi:10.1038/s43856-025-01158-1.

[17] DS@GT Team, Notebook for the erisk lab at CLEF 2025: Prompt engineering for conversational depression detection, in: Proceedings of the Working Notes of CLEF 2025, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[18] SINAI-UJA Team, Notebook for the erisk lab at CLEF 2025: SINAI-UJA approaches to depression detection, in: Proceedings of the Working Notes of CLEF 2025, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[19] DeepSeek-AI, DeepSeek-V3 technical report, arXiv preprint arXiv:2412.19437 (2024).

[20] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[21] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, Advances in Processing Systems 33 (2020) 5776–5788.

[22] A. Perez, J. Parapar, X. Wang, F. Crestani, TalkDep: Clinically grounded LLM personas for conversation-centric depression simulation, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2025. doi:10.1145/3746252.3761617.

[23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: Proceedings of the 10th International Conference on Learning Representations (ICLR), 2022.

[24] Y. W. Leow, et al., Leveraging large language models for cost-effective multilingual depression detection and severity assessment, arXiv preprint (2025).

[25] L. Weidinger, J. Uesato, M. Rauh, et al., Taxonomy of risks posed by language models, Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (2022) 214–229. doi:10.1145/3531146.3533088.