Expert Decomposition for Mental-Health Dialogue Analysis: Zero-Shot Psychometric Inference and Response Selection with Pool of Experts

Patrizio Bellan[1], Piergiorgio Maruotti[1], Leonardo Sanna[1] and Mauro Dragoni[1]

[1]*Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy*

Abstract

This paper presents our submission to MentalRiskES 2026, a shared task on mental-health analysis in Spanish therapeutic conversations. We investigate whether the Pool of Experts (PoE), a prompt-based identity-aware multi-agent architecture, can support zero-shot reasoning for two clinically sensitive tasks: questionnaire-based symptom estimation and therapist response selection. In PoE, a Project Manager identifies task-relevant expertise fields, multiple Expert Agents produce independent judgments, and a Final Decision Maker aggregates their outputs into the official prediction. Before the shared-task evaluation, we performed a preliminary grid search on external empathy and Theory-of-Mind datasets to select the backbone model and the profile-description framework for the main personality-aware configuration. Based on this selection, we submitted three runs: a Flow-Theory-based personality-aware PoE configuration; a no-personality run where agents still receive generic role-specific profile descriptions, but these descriptions are not generated from any psychological or design framework; and a no-description control run where no generated profile is appended to the agents, which are conditioned only by their functional role (expertise field), task instructions and context.

The results show a clear contrast between the two tasks. In Task 1, our best run ranked first overall, suggesting that expert decomposition can be effective for structured psychometric inference when the output space is organized around explicit questionnaire items. In Task 2, performance was weaker, indicating that identifying the most therapeutically appropriate response may require a finer-grained modeling of dialogue context and therapeutic adequacy.

Overall, our findings suggest that PoE is promising as an auditable support framework for structured questionnaire-oriented analysis.

The results, analyses, and code associated with this work are available at https://github.com/patriziobellan86/PoE-MentalRiskES-IberLEF-2026.

Keywords

Mental Symptom Detection, Pool of Experts, Multi-Agents System, Therapeutic Conversations Prediction, Zero-Shot, Psychometric Assessment Prediction, Agent Personality

1. Introduction

The rapid growth of Large Language Models (LLMs) has enabled the development of persona agents capable of adopting structured roles and simulating psychological perspectives [1, 2]. These systems are increasingly deployed in domains such as digital health, coaching, education, and therapeutic support, where they must interact with human reasoning and emotional attunement [3]. As these applications expand, evaluating perspective-taking capacity has become critical. For example, in traditional evaluations of empathy and theory of mind, LLMs have predominantly focused on single-agent architectures [4], leaving the ability of multi-agent systems grounded in psychological frameworks to demonstrate these capacities largely unexplored [5]. Multi-agent systems offer a promising alternative by simulating collaborative processes [6].

Mental health disorders represent a growing global challenge, with increasing demand for accessible and scalable support tools. Therapeutic conversations constitute a rich source of clinically relevant information, yet their systematic analysis remains largely manual, time-consuming, and dependent

IberLEF 2026, September 2026, León, Spain

pbellan@fbk.eu (P. Bellan); pmaruotti@fbk.eu (P. Maruotti); lsanna@fbk.eu (L. Sanna); dragoni@fbk.eu (M. Dragoni)

0000-0002-2971-1872 (P. Bellan); 0009-0005-8626-2590 (P. Maruotti); 0000-0003-3021-6606 (L. Sanna); 0000-0003-0380-6571 (M. Dragoni)

on expert interpretation. LLMs are increasingly being explored as components of systems for mental-health-related dialogue analysis, education, and therapist decision support. Their ability to process conversational language, summarize clinically relevant content, and reason over social and affective cues makes them potentially useful for psychotherapy-related workflows. At the same time, mental health remains a high-stakes domain where errors may affect how distress is interpreted, how risk is prioritized, and how therapeutic responses are evaluated.

These concerns are especially relevant for shared evaluation campaign *MentalRiskES 2026*. The competition focuses on real therapeutic conversations between patients and therapists, curated by domain experts. It evaluates systems on two related forms of mental-health reasoning: estimating questionnaire-based psychological indicators and selecting appropriate therapist responses. This combination makes the task relevant not only for symptom-oriented dialogue understanding, but also for the broader problem of supporting therapeutic decision-making in a controlled benchmark setting. The shared task includes two complementary subtasks. Task 1 focuses on structured psychometric inference: systems must predict item-level responses for three questionnaires, namely GAD-7, PHQ-9, and CompACT-10. This requires translating dialogue evidence into questionnaire scores that approximate clinically meaningful constructs such as anxiety, depressive symptoms, and psychological flexibility. Task 2 focuses on therapist response selection: systems must choose the most appropriate response among three candidate therapist interventions, taking into account the dialogue context and the therapeutic adequacy of each option. Together, the two tasks evaluate whether a system can both infer patient-side psychological information and support therapist-side response assessment.

We participated in MentalRiskES 2026 as the *FBKillers* team. Our submission explores whether a zero-shot personality-aware multi-agent architecture can provide useful structure for these tasks without task-specific fine-tuning. In particular, we adopt the Pool of Experts (PoE) framework [7], a multi-agent architecture where agents are equipped with identity description to simulate real person. In PoE, a *Project Manager* (PM) agent identifies task-relevant expertise fields, multiple *Expert Agents* (EAs) reason independently, and a *Final Decision Maker* (FDM) agent produces the final prediction. Interestingly, in configuring PoE for MentalRiskES evaluation campaign, we used only the task descriptions and instructions provided on the official shared-task page as contextual and task-level instructions for the system, without adding task-specific supervision or manually engineered examples.

Before the official MentalRiskES evaluation, we conducted a preliminary grid search to select the back-bone LLM for the agents and the most performing profile-description framework used to generate identity profiles of agents. This search was performed on proxy datasets on empathy and Theory-of-Mind. Based on this preliminary evaluation, we selected the top performing model–framework pair and submitted three configurations: (Run 0) a PoE run using *Flow Theory*; (Run 1) a no-personality run where agents still receive generic role-specific profile descriptions, but these descriptions are not generated from any psychological or design framework; (Run 2) and a no-description control run where no generated profile is appended to the agents, which are conditioned only by their functional role, expertise field, task context, and output instructions.

Our results show a strong contrast between the two tasks. In Task 1, the best FBKillers run ranked first overall, achieving the lowest combined MAE among all submitted systems. The system performed particularly well in the structured questionnaire prediction setting, suggesting that expert decomposition can be useful when the output space is organized around explicit psychometric items. In Task 2, however, performance was substantially weaker, indicating that selecting the most therapeutically appropriate response remains more difficult for the current architecture.

After the competition, the organizers released the official gold labels, allowing us to perform additional post-hoc analyses of expert behavior and aggregation. These analyses show that the FDM agent can recover some majority-vote errors, but also that it frequently fails to select correct minority answers already present in the expert pool. This suggests that the main limitation is not only the quality of individual expert predictions, but also the decision mechanism used to integrate them. In all official submissions, the final answer was produced by the FDM. Comparisons between the FDM and majority vote are therefore performed only as post-hoc analyses over internal agent traces, not as submitted competition runs.

The main contributions of this paper are as follows. First, we demonstrate the potential of PoE as zero-shot architecture for solving mental-health dialogue analysis. Second, we describe a preliminary grid-search procedure for selecting the model and profile-description framework used in the official evaluation. Third, we provide a post-hoc analysis of expert aggregation, comparing FDM behavior against majority vote and examining expert-level performance to discuss implications for future multi-agent systems in clinically sensitive NLP settings.

2. Related Work

This section reviews three lines of research that are directly relevant to our work. First, we discuss studies on personality- and persona-based conversational agents, with particular attention to how profile descriptions can shape LLM behavior. Second, we review multi-agent LLM systems, focusing on coordination, expert decomposition, and aggregation mechanisms. Third, we examine recent work on LLM-based systems for mental-health analysis, patient simulation, and the evaluation of therapeutic conversations. Together, these areas provide the methodological and application-oriented context for our use of a personality-aware Pool of Experts architecture in the MentalRiskES 2026 shared task.

2.1. Agent Personality

Conversational agents have frequently incorporated psychological frameworks, especially the Big Five, to support coherent interaction styles and enable forms of personalization [8, 9]. In applied settings such as healthcare, persona-based design has been shown to affect users' perceptions of empathy and trust, as well as task-related outcomes [10, 11]. More recently, both prompt-based and data-driven strategies have been explored to increase contextual sensitivity and diversify the conversational behavior of LLM-based agents [12, 13]. A growing body of work has also examined persona diversity at scale, evaluating whether synthetic personas preserve stable traits and how Big Five dimensions or social-identity cues influence downstream model behavior [14, 15]. However, these studies often concentrate on a single personality framework or on a limited set of evaluation tasks. Personality- and identity-based prompting further suggests that profile descriptions can exert a *propagated* influence across agent behavior. Big Five traits and upstream persona definitions can generate consistent behavioral patterns and affect downstream agents [15, 16], including longer-term effects on task selection and interaction dynamics [17]. These findings motivate our use of psychologically grounded identities as a controllable inductive bias for both Expert Agents and the coordination layer. Beyond explicit trait prompting, recent work has investigated how LLMs can *generate* profile descriptions and produce persona-consistent outputs [16], as well as how automatically generated expert roles can increase diversity and improve factual reliability in multi-agent reasoning [18, 19].

2.2. Multi-Agent Systems

Previous research on multi-agent systems shows that coordination mechanisms can strongly influence group behavior. Examples include debate-style judges, proactive controllers, and hierarchical leaders that guide planning, align intermediate outputs, and synthesize the contributions of multiple agents [20, 21, 22].

Control-loop architectures and multi-expert pipelines further demonstrate that monitoring, critique, and synthesis components can redirect or refine the reasoning of other agents [23, 19, 24]. Simpler ensemble methods, such as majority voting, instead rely on protocol-level aggregation without an explicit reasoning-oriented coordinator. In contrast, our work conditions the identity of the aggregation agent itself: the `Final Decision Maker` agent is instantiated with, or without, psychologically grounded profile descriptions.

2.3. Mental-Health Systems and Therapeutic Conversation Evaluation

The application of multi-agent LLM architectures to mental-health tasks has received increasing attention in recent years. DSM5AgentFlow is a multi-agent workflow in which a therapist agent administers DSM-5 diagnostic questionnaires to a simulated client agent, while a diagnostician agent produces transparent, step-by-step disorder predictions grounded in DSM-5 criteria [1]. This work shares with ours the use of specialized agents with distinct roles and suggests that structured multi-agent orchestration can support both accuracy and explainability in clinically oriented evaluation settings. Similarly, the system proposed in [25] comprises a four-agent architecture for depression assessment from clinical interview transcripts, combining qualitative symptom identification, iterative self-refinement, embedding-based few-shot score prediction, and meta-review integration. The system achieved human-level accuracy on PHQ-8 severity prediction without fine-tuning, supporting the viability of zero-shot and few-shot multi-agent approaches for psychometric estimation from therapeutic text. More broadly, the work proposed in [26] demonstrate the utility of cooperative multi-agent LLM architectures for generating synthetic behavioral-health data from natural-language descriptions of user personas. Their work highlights how role-specialized agents coordinated by a supervisor can address complex and interdependent tasks that single-agent systems may handle less reliably, a design principle that also motivates our use of the Pool of Experts framework for structured psychometric inference and response selection.

A complementary line of research has explored the use of LLMs to simulate patients and evaluate therapeutic interactions. Client101 [27] is a web-based platform in which GPT-4-based chatbots simulate patients with depression and generalized anxiety disorder for psychotherapy training purposes. Using prompt engineering grounded in clinical vignettes and validated with licensed psychologists, the system produced conversations that mental-health professionals found largely realistic and clinically useful. Psycholinguistic analysis further indicated that several linguistic markers associated with depression and anxiety were preserved in the synthetic transcripts, although differences in thematic content and emotional range also highlighted the limits of static prompt-based simulation.

A related line of work has directly compared LLM-based chatbot responses with those produced by licensed therapists. In [28], the authors conducted a mixed-methods study in which 17 licensed therapists evaluated the responses of general-purpose chatbots to scripted mental-health scenarios. Their findings show that, although chatbots can display some elements associated with good therapeutic interaction, such as validation and reassurance, they also tend to overuse directive advice, ask too few open-ended questions, and perform poorly in crisis situations. This systematic characterization of the gap between chatbot and therapist behavior is directly relevant to our Task 2 findings. In our experiments, selecting the most therapeutically appropriate response proved substantially more difficult than estimating questionnaire-based psychological indicators, suggesting that the qualities distinguishing expert therapeutic responses remain difficult to operationalize in current LLM-based systems.

3. Shared Tasks on Early Mental Health Risk Detection in Spanish

MentalRiskES is a shared-task series organized within the Iberian Languages Evaluation Forum (IberLEF), focused on mental-health analysis in Spanish. Previous editions addressed risk detection from social-media data. The second edition, held at IberLEF 2024, proposed three subtasks: multiclass disorder detection, covering depression, anxiety, and no disorder; contextual risk-factor identification; and suicidal-ideation detection [29]. With 28 registered teams and 12 teams submitting results, participating systems relied predominantly on supervised fine-tuning of Spanish Transformer-based models such as RoBERTa and BETO. The third edition, held at IberLEF 2025, shifted the focus to addiction-risk detection, specifically gambling disorder, using a new corpus of Spanish social-media messages [30]. Across these editions, the dominant methodological paradigm was supervised learning on in-domain labeled data.

Within IberLEF 2026 [31], MentalRiskES 2026 [32], the edition in which our system participates, continues this line of shared evaluation campaigns while introducing a substantially different evaluation

setting, extending mental-health analysis in Spanish to a new dialogue-based clinical scenario. Rather than focusing on social-media posts, it is based on real therapeutic conversations in Spanish, curated and annotated by domain experts. The task setting requires systems to reason over patient–therapist interactions and to produce predictions without access to labeled training data for model fine-tuning. This shift makes the direct application of prior supervised approaches less suitable and motivates the exploration of zero-shot, instruction-based, and knowledge-guided architectures such as our Pool of Experts framework [33].

MentalRiskES 2026 consists of two tasks. Task 1 evaluates questionnaire-based psychological inference from therapeutic dialogue, while Task 2 evaluates therapist response selection. Together, these tasks assess two complementary dimensions of mental-health dialogue analysis: estimating patient-side psychological indicators and identifying therapist-side responses that are appropriate within the ongoing interaction.

Task 1: Early Symptom Detection. This task requires systems to estimate, after each patient turn, how the patient would respond to three validated psychometric instruments: the Generalized Anxiety Disorder scale (GAD-7) [34], the Patient Health Questionnaire (PHQ-9) [35], and the Comprehensive Assessment of Acceptance and Commitment Therapy processes (CompACT-10) [36]. For each questionnaire, systems must assign a value to every item according to the corresponding Likert scale. The task therefore requires mapping evidence from the therapeutic conversation onto structured questionnaire responses.

Task 2: Therapist Response Selection. This task requires systems to select, at each interaction step, the most appropriate therapist response among three candidate options. One candidate response is authored by a clinical expert, while the other two are automatically generated. The task therefore evaluates whether systems can identify the response that best fits the therapeutic context, considering aspects such as relevance, empathy, continuity with the dialogue, and clinical appropriateness.

4. The Pool of Experts Framework

The **Pool of Experts (PoE)** framework is a prompt-based multi-agent architecture originally introduced in [7]. Its central idea is to replace a single direct LLM prediction with a structured reasoning process involving multiple role-specialized agents. Rather than asking one model instance to solve the task in isolation, PoE decomposes the problem into three main operations: selecting relevant expertise, generating independent expert judgments, and aggregating these judgments into a final answer. The framework is inspired by the way interdisciplinary teams approach complex decisions: different specialists examine the same problem from different perspectives, and a final decision is produced after comparing their contributions.

In the PoE framework, an *agent* is an LLM instance conditioned by a role-specific system prompt. All agents can be instantiated from the same underlying LLM; therefore, their differences do not arise from different model parameters, fine-tuning, or task-specific training. Instead, diversity is introduced at the prompt level through three components: the agent's functional role, its assigned field of expertise, and, when enabled, a profile description generated according to a description framework.

PoE is composed of four main agent roles: the Psychologist Agent, the Project Manager Agent, the Expert Agents, and the Final Decision Maker.

The **Psychologist Agent** (PA) is responsible for generating profile descriptions for the other agents. In configurations where profile-based prompting is enabled, the PA receives the name of a description framework, e.g., Big Five, together with the context and the task instruction, and uses them to create role-specific descriptions. These descriptions are not manually written for each expert. Instead, they are generated systematically by the PA. For example, the PA may generate a profile for a Project Manager emphasizing coordination and planning, or a profile for an Expert Agent emphasizing a specific domain of expertise. The PA does not solve the final task directly. Its purpose is to introduce controlled profile

variation into the system. In this study, this mechanism is used to test whether different prompt-level descriptions can influence zero-shot task behavior. We do not interpret the generated profiles as real psychological states of the model; they are prompt-level constraints intended to shape the style and focus of the agent's reasoning.

The **Project Manager Agent** (PM) Agent acts as selector of expertise. Given the task description and the task context, the PM identifies which kinds of experts should participate in the reasoning process. In the MentalRiskES setting, these expertise fields may include, for example, clinical psychology and mental health assessments. The PM is therefore responsible for translating the task into an expert-pool structure. This step is important because it makes the composition of the multi-agent system task-dependent. Rather than using a fixed set of experts for every possible problem, PoE allows the pool to be adapted to the nature of the task.

Expert Agents (EAs) are the main producers of candidate answers. Each EA receives the same task instance, but reasons from the perspective of its assigned expertise field. The agents operate independently: they do not deliberate with each other, revise each other's outputs, or negotiate a shared answer before aggregation. This design makes their outputs analytically useful, because disagreement among experts can be inspected after inference.

The **Final Decision Maker** (FDM) agent receives the set of Expert Agent outputs and produces the final prediction. The FDM is prompted to evaluate the candidate answers provided by the EAs by considering their consistency, confidence scores, reasoning steps and justifications. This makes the FDM different from simple majority voting. Majority vote selects the most frequent answer among the experts. The FDM, instead, can in principle select a minority answer if it is better justified or more coherent with the task context.

4.1. Identity-Aware Agent

A distinctive component of PoE is the possibility of equipping agents with generated profile descriptions. A profile description specifies how the agent should approach the task, communicate, reason under uncertainty, and prioritize evidence. These profiles are produced according to a selected description framework. Such frameworks may emphasize personality traits, cognitive processes, motivational dynamics, or user-oriented perspectives.

In this paper, profile-based variation is treated as an experimental prompting intervention. It does not imply that the model has a stable personality, nor that the generated profile corresponds to an actual internal cognitive structure. Rather, profile descriptions are used to introduce controlled differences among agents instantiated from the same LLM. This allows us to study whether expert decomposition combined with profile variation affects performance, disagreement, and aggregation behavior.

The official submissions include three configurations. In the first configuration, agents are generated using **Flow Theory** as the selected profile-description framework. This represents the main PoE configuration chosen after preliminary model–framework selection. In the second configuration, referred to as **no-personality**, agents are described without grounding their profiles in a specific psychological framework. In the third configuration, referred to as **no-description**, agents are assigned expertise roles but are not given additional personality descriptions. These controls allow us to distinguish between the effect of framework-based profile generation and generic role description.

Thus, we distinguish between three levels of profile conditioning. In the *Flow-Theory* configuration, the Psychologist Agent generates role-specific profile descriptions grounded in Flow Theory; these descriptions are incorporated into the system prompts of the Psychologist agent, Project Manager agent, Expert Agents, and Final Decision Maker agent. In the *no-personality* configuration, the Psychologist Agent still generates role-specific profile descriptions, but no psychological or design framework is provided as a guiding schema. Thus, agents receive generic descriptions of how they should approach the task, communicate, reason under uncertainty, and use their expertise, but these descriptions are not grounded in a named personality, cognitive, or design framework. In the *no-description* configuration, this profile-generation step is disabled: agents are conditioned only by their functional role, assigned expertise field, task context, and output format instructions. Therefore, no-description does mean that

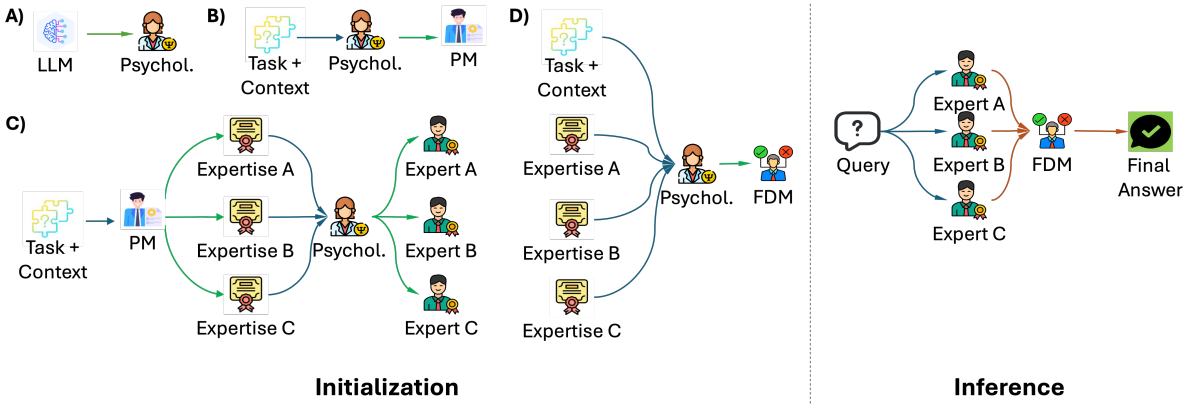


Figure 1: The figure, taken from [7], shows an overview of the PoE framework.

agents has a role and no additional generated profile or personality-style description is included in the agent system prompt.

4.2. Execution Pipeline

PoE execution consists of two phases: initialization and inference.

During initialization (Fig.1 Initialization), the system constructs the agent pool. First, the Psychologist Agent is instantiated (Fig.1 A). Then, depending on the configuration, the PA generates role-specific descriptions for the Project Manager (Fig.1 B), the Expert Agents (Fig.1 C), and the Final Decision Maker(Fig.1 D). The Project Manager analyzes the task and defines the expertise fields needed for the pool. For each selected expertise field, an Expert Agent is instantiated. Finally, the FDM is instantiated as the aggregation component.

This phase determines the structure of the reasoning system before any individual test instance is processed. In other words, initialization defines *who* the agents are, *which expertise fields* are represented, and *which description strategy* is used to condition their behavior.

During inference (Fig.1 Inference), each input instance is sent to all Expert Agents in the pool. Each EA independently produces an answer. The set of expert answers is then passed to the FDM, which generates the final prediction. In parallel, because the Expert Agent outputs are stored, we can compute majority vote after inference and compare it with the FDM.

This separation between expert prediction and final aggregation is central to PoE. It allows us to analyze not only whether the final answer is correct, but also how the system arrived at it. For example, we can identify cases where most experts were wrong but the FDM selected the correct answer, cases where at least one expert was correct but the FDM failed to use that information, and cases where majority vote would have outperformed the FDM.

5. Grid Search

Before the official MentalRiskES evaluation, we conducted a preliminary grid search to select the LLM and the profile-description framework to be used in the main PoE configuration. The goal of this step was not to tune the system on MentalRiskES data, but to identify a robust model–framework combination on external tasks related to empathy and social reasoning. For this reason, the grid search was performed on small stratified subsets of external empathy and Theory-of-Mind benchmarks.

The grid search varied two main dimensions: the backbone LLM and the description framework used to generate personality-aware agent profiles. Each tested configuration instantiated the full PoE pipeline and was compared under the same evaluation protocol. The selected configuration was then used as the main personality-aware run in the official competition. In addition, we included two ablation controls:

no-personality, where agents were described without using a specific profile-description framework, and *no-description*, where agents were assigned expertise roles without additional personality descriptions.

The description frameworks are used to guide the generation of the personality of the agents. These are used in the grid search where we selected a representative set of them to cover different ways of specifying agent behavior. Some frameworks emphasize cognitive or psychological mechanisms, while others describe users, goals, or interaction patterns from a design-oriented perspective. The full list of description framework we used during grid search is presented in Table 1).

Table 1
Description frameworks.

Framework	Group	Role in the grid search
Flow Theory [37]	Motivation theory	Models agents through focus, intrinsic motivation, and engagement with challenging tasks. This framework was selected for the main personality-aware official run.
Dual-Process Theory [38]	Neuro-cognitive theory	Models reasoning as the interaction between fast, intuitive processes and slower, deliberative processes.
Cognitive Load Theory [39]	Learning theory	Models agents in terms of cognitive resource management, information load, and structured processing.
Freudian Psychoanalysis [40]	Mental functioning	Uses a psychodynamic perspective to generate profiles shaped by internal drives, conflicts, and interpretive tendencies.
User-Centered Design [41]	Methodology	Represents agents through needs, goals, constraints, and interaction-oriented reasoning.
User Design Persona [42]	Methodology	Generates profile descriptions inspired by persona-based design, emphasizing goals, behaviors, and task-oriented characteristics.
no-personality	Control	Control condition in which agents are described without following a specific psychological or design framework.
no-description	Control	Control condition in which agents are assigned expertise roles without additional personality descriptions.

The grid search also evaluated multiple LLMs from different model families. The selected models were chosen to cover variation in size, architecture, and expected inference cost. Table 2 summarizes the models considered in the preliminary evaluation.

Table 2
Backbone LLMs.

Family	Model	Parameters	Architecture
Gemma	Gemma 3	4B	Dense
Gemma	Gemma 3	12B	Dense
Gemma	Gemma 3	27B	Dense
Llama	Llama 3.1	8B	Dense
Llama	Llama 3.2	3B	Dense
Llama	Llama 3.3	70B	Dense
Llama	Llama 4 Scout	109B	MoE
Mistral	Mistral Nemo	12B	Dense
Nova	Nova Micro v1	N/A	Dense

The grid search was conducted on proxy datasets to select a model–framework configuration using

tasks that were conceptually related to the target domain, but independent from the official evaluation set. We therefore used benchmarks covering two broad dimensions: empathy-related language understanding and Theory-of-Mind (ToM) reasoning. These dimensions were selected because both are relevant to therapeutic dialogue analysis: the former concerns the recognition of emotional and supportive communicative signals, while the latter concerns the ability to reason about beliefs, intentions, emotions, and perspectives. For empathy-oriented evaluation, we used three datasets. *EDOS* [43] contains dialogue data annotated with fine-grained emotion labels. *IEMPATHIZE* [44] focuses on empathy direction, distinguishing whether an utterance expresses seeking empathy, providing empathy, or neither. *EmpatheticDialogues* [45] contains conversations grounded in emotional situations and labeled across a broad set of emotion categories. Together, these datasets allow the grid search to test whether a configuration can handle affective and interpersonal information in dialogue-like settings. For Theory-of-Mind-oriented evaluation, we used three datasets. *Hi-ToM* [46] includes story-based questions designed to probe different levels of mental-state reasoning. *ToMATO* [47] evaluates reasoning over multiple mental-state categories, including beliefs, desires, intentions, emotions, and knowledge. *OpenToM* [48] contains narrative scenarios requiring models to answer questions involving location tracking, multi-hop inference, and attitude attribution. These datasets were used to assess whether the model–framework combinations could support perspective-taking and inference over implicit mental states. Due to computational constraints, the grid search was performed on a small random stratified subset of 10 samples for each dataset.

Table 3 reports the seven best configurations identified in the preliminary grid search. For each model–framework combination, we evaluated the full PoE pipeline on the datasets described above. Performance was computed by first evaluating each configuration on every sampled dataset and then averaging the resulting scores across datasets, so that no single benchmark determined the final ranking. The results motivated us to adopt the FDM aggregation strategy over the majority vote one. Interestingly, the relatively small differences among the top configurations suggest that no one description framework is generally superior across mental-health NLP tasks.

Based on this preliminary comparison, we used *Flow Theory* [37] using `openai/gpt-4o-mini` as back-bone model for the main PoE run submitted to the official evaluation. The two control conditions, *no-personality* and *no-description*, were retained as additional official runs to assess whether performance depended on personality-based profile generation or primarily on expert-role decomposition.

Table 3
Grid search: top seven configurations.

Rank	Model	Description Framework	Accuracy FDM	Accuracy MV
1	`openai/gpt-4o-mini`	*Flow Theory* [37]	65.00 %	60.00 %
2	`openai/gpt-4o-mini`	*Cognitive Load Theory* [39]	63.33 %	61.67 %
3	`openai/gpt-4o-mini`	*Freudian Psychoanalysis* [40]	63.33 %	61.67 %
4	`openai/gpt-4o-mini`	*User-Centered Design* [41]	63.33 %	61.67 %
5	`openai/gpt-4o-mini`	*Dual-Process Theory* [38]	61.67 %	61.67 %
6	`google/gemma-3-27b-it`	*Cognitive Load Theory* [39]	61.67 %	60.00 %
7	`openai/gpt-4o-mini`	*User Design Persona* [42]	61.67 %	58.33 %

6. Experimental Settings

We adapted PoE to the two MentalRiskES tasks while keeping the same general architecture. In all cases, we used only the task descriptions and instructions provided on the official shared-task page as contextual and task-level instructions. We did not use task-specific fine-tuning, manually engineered examples from the test set, or additional supervised training.

At each evaluation step, the dialogue context was constructed cumulatively. More specifically, for a given round, the input instance provided to the agents consisted of the full patient–therapist

conversation available up to that point. Therefore, the model did not receive only the most recent utterance pair, nor a fixed-size sliding window of previous turns.

For **Task 1**, the goal is to predict item-level responses for three questionnaires: GAD-7, PHQ-9, and CompACT-10. We initialize PoE once for each questionnaire subtask. We deliberately did not instantiate PoE for each questionnaire item. An item-specific design could plausibly improve performance, because each item has its own semantic focus and scoring anchors. For example, an item about sleep disturbance may benefit from a different expert composition than an item about guilt, worry, avoidance, or psychological flexibility. However, creating one pool per item would multiply the number of agent calls, substantially increasing computation time, cost, and associated emissions. The adopted design therefore reflects a trade-off between specialization and computational sustainability.

For **Task 2**, the goal is to select the most appropriate therapist response among three candidate options. We instantiated PoE for this response-selection task. Each Expert Agent evaluated the candidate responses in light of the dialogue context and selected one option. The FDM then produced the official submitted prediction.

The same architecture thus supports two different forms of mental-health reasoning using only different initialization.

7. Results and Discussion

The official Task 1 results indicate that questionnaire-based inference is the strongest setting for our system. The best FBKillers run ranked first overall on the official leaderboard. This suggests that prompt-based expert decomposition can be effective when the task is organized around explicit assessment dimensions, as in GAD-7, PHQ-9, and CompACT-10. The comparison among the three submitted configurations is particularly informative (see Table 4).

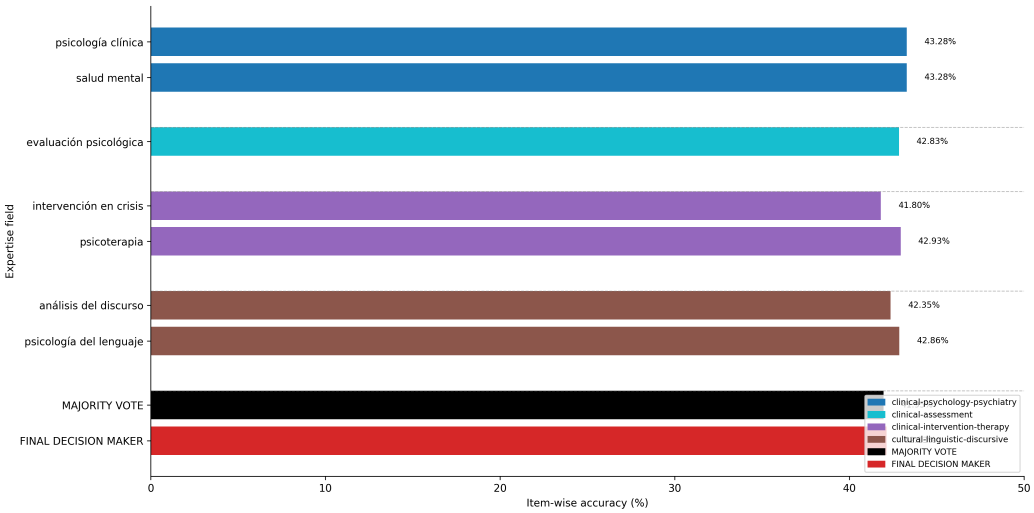
Table 4
team runs and official ranks.

Rank	Run	configuration
1	2	control setting *no-description*
5	1	control setting *no-personality*
7	0	personality-aware using *Flow Theory*

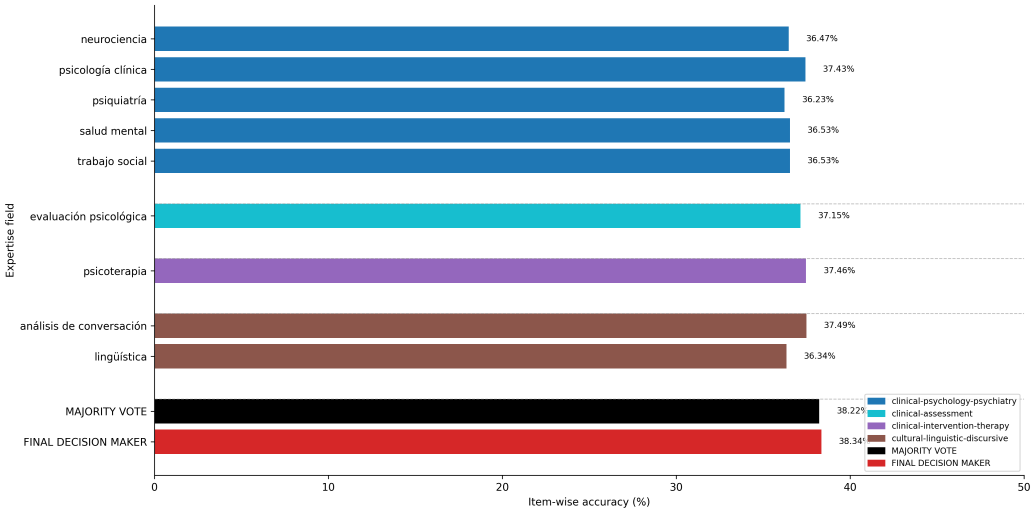
Results show that our Run 2 performed the best, followed by Run 1 and Run 0. Since the no-description configuration removes the generated profile layer while preserving role and expertise specialization, this result suggests that, for questionnaire-based inference, the main useful component may be expert-role decomposition rather than profile-based behavioral conditioning. This result does not imply that personality descriptions are generally ineffective. Personality framing can enrich agent diversity, but it may also introduce stylistic or interpretive variability that is not always useful for mapping dialogue evidence onto ordinal questionnaire scores. The pattern is especially visible for CompACT-10, where Run 2 showed the strongest gains. Since CompACT-10 targets psychological flexibility, acceptance, and committed action, reducing prompt complexity may have helped the model focus more directly on item semantics and dialogue evidence.

After the release of the gold labels, we reconstructed additional diagnostic signals from the internal Expert Agent outputs, including Majority Vote (MV) and expert-level correctness. Figure 2 provides the results for each EA. We grouped the agents based on their main specialization group. In general the group perform almost the same, with minimal variation among the three questionnaires. The only exception regards the CompACT-10 questionnaire, where we see more variability among them. This indicate that this sub-task is more challenging than the other two. Furthermore, the two aggregation strategies (Majority Vote and Final Decision Maker agent) show almost no difference.

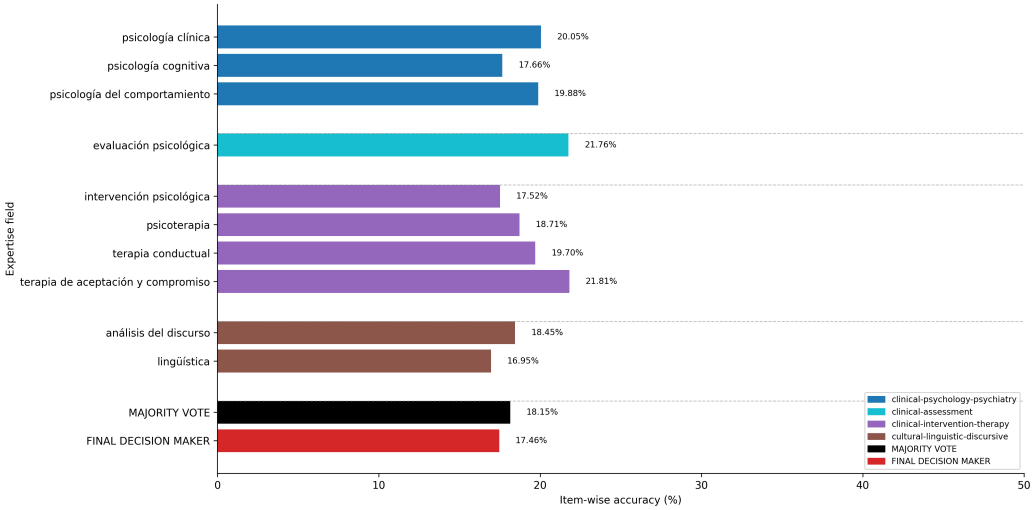
Post-hoc analyses allows us to analyze whether the FDM corrects MV errors, whether it succeeds despite partial expert disagreement, and whether correct predictions are present in the expert pool but not selected by the FDM. We use five diagnostic categories: *FDM recovers MV error*, where the FDM is



(a) GAD-7.



(b) PHQ-9.



(c) CompACT-10.

Figure 2: Task 1 Run 0, item-wise accuracy of PoE agents on the three questionnaires.

correct and MV is wrong; *FDM recovers all agents wrong*, where the FDM is correct despite all Expert Agents being wrong; *FDM recovers partial agent errors*, where the FDM is correct while at least one Expert Agent is wrong; *Any agent > FDM*, where the FDM is wrong while at least one Expert Agent is correct; and *MV > FDM*, where MV is correct while the FDM is wrong. Table 5 reports the resulting counts and item-level accuracies.

Table 5
Post-hoc Task 1 error analysis.

Subtask	Run	N Items	FDM Acc.	MV Acc.	Avg. Agent Acc.	FDM Recovers MV	FDM Recovers All Wrong	FDM Recovers Partial	Any Agent > FDM	MV > FDM
GAD-7	0	3976	42.13%	41.95%	42.76%	93	0	702	814	86
GAD-7	1	3976	42.43%	42.23%	42.65%	123	0	676	745	115
GAD-7	2	3976	43.13%	43.01%	43.07%	96	0	712	743	91
PHQ-9	0	5112	38.34%	38.22%	36.85%	169	0	1298	1067	163
PHQ-9	1	5112	38.81%	37.68%	36.88%	218	0	1373	1136	160
PHQ-9	2	5112	40.00%	38.11%	37.32%	225	0	1408	1077	128
CompACT-10	0	5680	17.46%	18.15%	19.25%	146	1	890	1939	185
CompACT-10	1	5680	20.60%	22.24%	22.78%	204	1	1070	2197	297
CompACT-10	2	5680	30.88%	31.65%	29.27%	214	0	1516	1630	258

Post-hoc analyses show different aggregation dynamics across questionnaires. For GAD-7, FDM and MV are very close, suggesting that the expert pool often converges on similar item-level predictions. For PHQ-9, the FDM is more useful: it consistently improves over MV and the average Expert Agent, indicating that the final aggregation layer can sometimes select better-supported predictions from heterogeneous expert outputs. CompACT-10 presents the opposite pattern: MV and individual Expert Agents are often competitive with, or better than, the FDM, suggesting that the aggregation mechanism can degrade the signal already present in the pool of EAs answers. The category *FDM recovers all agents wrong* is almost absent. This indicates that the FDM rarely generates an independent corrective signal when the full expert pool fails. Most successful recoveries occur when at least one Expert Agent already provides useful evidence. Overall, the analysis suggests that future improvements should focus on aggregation, including confidence calibration, item-specific evidence weighting, and questionnaire-specific decision rules.

Table 6
Team runs and official ranks.

Rank	Run	configuration
32	1	control setting *no-personality*
33	0	personality-aware using *Flow Theory*
38	2	control setting *no-description*

The results of Task 2 presents a different pattern. Indeed, therapist response selection is considerably more challenging for our system, with FBKillers runs ranking in the lower part of the leaderboard (see Table 6). Here, our Run 1 (*no-personality*) was the best FBKillers configuration, Run 0 performed similarly, and Run 2 (*no-description*) was the weakest. This suggests that therapist response selection may benefit from some descriptive role framing, but, maybe not necessarily using Flow Theory. Simply assigning expertise labels may leave too much of the therapeutic evaluation implicit, whereas a moderate level of agent description may provide useful behavioral structure. The analysis of the Expert Agents performance indicates that multi-agent decomposition alone is insufficient for this task. Selecting the best therapeutic response likely requires more explicit evaluation criteria, such as emotional validation, non-judgmental stance, relevance, safety, and avoidance of premature advice.

Table 7
Post-hoc Task 2 error analysis at item level.

Run	N Items	FDM Acc.	MV Acc.	Avg. Agent Acc.	FDM Recovers MV	FDM Recovers All Wrong	FDM Recovers Partial	Any Agent > FDM	MV > FDM
0	568	20.07%	19.54%	20.14%	6	0	71	99	3
1	568	20.25%	18.66%	19.98%	12	0	74	103	3
2	558	19.35%	18.10%	20.41%	9	0	61	96	2

We then performed the post-hoc analysis. Figure 3 shows that differences among EAs, MV, and FDM are small, suggesting that the limitation concerns both aggregation and the suitability of the current

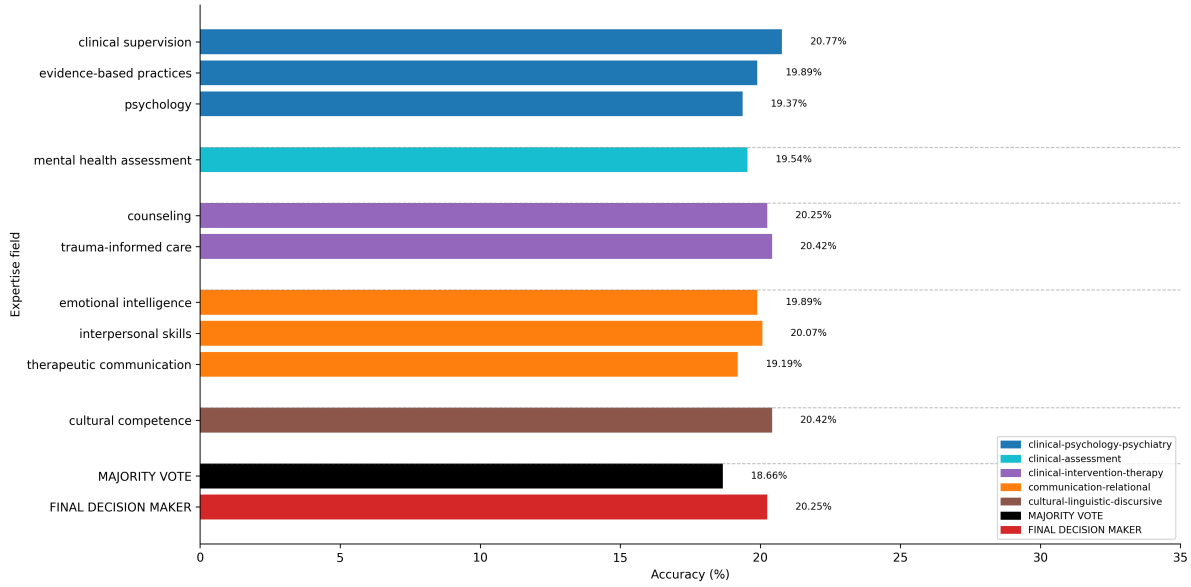


Figure 3: Representative Task 2 performance for the best FBKillers configuration. The plot compares Expert Agent accuracy with reconstructed Majority Vote and the official Final Decision Maker output.

expert prompts. We also run for this task an error analysis using the same diagnostic categories adopted for Task 1 (see Figure.7). This analysis shows that the FDM provides only a limited advantage over MV. The number of MV errors recovered by the FDM is small, while cases in which at least one Expert Agent is correct but the FDM is wrong are more frequent. This indicates that correct minority signals are sometimes present in the pool, but the current FDM does not reliably identify them. The analysis shows that the FDM only slightly improves over MV in all three runs. These differences suggest that the FDM provides some benefit over simple consensus, but the gain is modest and does not substantially change the overall difficulty of the task. The number of cases in which the FDM recovers an MV error is also limited. This indicates that the FDM occasionally selects the correct response when the majority of Expert Agents converges on a wrong option, but this recovery behavior is not frequent enough to compensate for the broader weakness of the system on therapist response selection. As in Task 1, the category *FDM recovers all agents wrong* is always zero. This means that the FDM never produces a correct answer when all Expert Agents are wrong. Correctly, the FDM does not generate an independent answer from the EAs. The most informative result concerns the *Any agent > FDM* category. These counts suggests that the expert pool often contains a correct minority signal, but the FDM is not sufficiently effective at identifying and prioritizing it. Finally, MV rarely succeeds when the FDM fails. Consequently, the main aggregation problem is not that majority voting is systematically better than the FDM. Rather, both aggregation strategies struggle because correct answers are often isolated within minority expert outputs and are not reliably selected by either consensus or final decision reasoning. Error analysis confirms that therapist response selection is substantially more challenging for the current PoE configuration than structured questionnaire prediction. The FDM provides a small advantage over MV, but it does not reliably exploit correct minority expert predictions. This points to a need for more explicit therapeutic response evaluation criteria, better dialogue-state tracking, and aggregation mechanisms that can assess the quality of expert justifications rather than relying mainly on answer distribution or weak synthesis.

Overall, the main bottleneck of our system is not simply majority voting versus FDM aggregation, but the lack of a sufficiently robust mechanism for recognizing the best-supported therapeutic response among competing expert outputs. Future work should work on this direction. The contrast between the two tasks is central. PoE appears effective for structured psychometric inference, but is, however, less effective for therapist response selection, where appropriateness depends on the evolving therapeutic relationship and on subtle pragmatic judgments. This distinction has practical implications: PoE-like

systems may be useful as assistive tools for monitoring questionnaire-related indicators or flagging changes in patient state, but they should not be interpreted as autonomous systems for therapeutic decision-making. From a clinical perspective, the Task 1 results suggest that zero-shot multi-agent systems may help reduce the administrative burden of repeated questionnaire administration and support clinicians in tracking patient progress between sessions. However, the Task 2 results highlight the limits of current LLM-based systems in modeling the relational and dynamic nature of psychotherapy. Therapeutic responses do not merely reflect symptom content; they shape the interaction itself. This reinforces the need for human oversight, transparent uncertainty reporting, and careful ethical constraints when applying such systems in mental-health contexts.

8. Conclusion

This paper presented the FBKillers submission to MentalRiskES 2026, based on the Pool of Experts framework. We evaluated a prompt-based, personality-aware multi-agent architecture on two Spanish mental-health dialogue tasks: questionnaire-oriented psychometric inference and therapist response selection. The system used only the official task descriptions as task-level instructions and did not rely on task-specific fine-tuning.

The results show that PoE is more effective for structured questionnaire prediction than for therapist response selection. In Task 1, our best run ranked first overall, suggesting that expert decomposition can be useful when the target output is organized around explicit questionnaire items. In Task 2, however, the system obtained substantially weaker results, indicating that selecting the most appropriate therapist response requires more precise modeling of therapeutic adequacy than our current architecture provides.

The post-hoc analyses further show that aggregation is a central challenge. The Final Decision Maker can recover some errors, but it also fails in many cases where correct answers are already present among the expert agents. This suggests that future work should focus on developing more reliable and transparent aggregation mechanisms.

Overall, PoE appears more promising for structured, auditable, questionnaire-oriented support than for autonomous therapeutic response selection. Its most realistic application is therefore as a transparent assistant for clinicians or researchers, rather than as an independent therapeutic agent.

Declaration on Generative AI

During the preparation of this work, the authors used the generative AI tool OpenAI chatGPT-5.2 to support the proofreading and linguistic refinement of the manuscript. All content was carefully reviewed, verified, and approved by the authors.

References

[1] M. C. Ozgun, J. Pei, K. Hindriks, L. Donatelli, Q. Liu, J. Wang, Trustworthy AI psychotherapy: multi-agent LLM workflow for counseling and explainable mental disorder diagnosis, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, 2025, pp. 2263–2272.

[2] J.-t. Huang, M. H. Lam, E. J. Li, S. Ren, W. Wang, W. Jiao, Z. Tu, M. R. Lyu, Apathetic or empathetic? Evaluating LLMs' emotional alignments with humans, Advances in Neural Information Processing Systems 37 (2024) 97053–97087.

[3] V. Sorin, D. Brin, Y. Barash, E. Konen, A. Charney, G. Nadkarni, E. Klang, Large language models and empathy: systematic review, Journal of medical Internet research 26 (2024) e52597.

[4] K. Gandhi, J.-P. Fränken, T. Gerstenberg, N. Goodman, Understanding social reasoning in language models with language models, Advances in Neural Information Processing Systems 36 (2023) 13518–13529.

[5] N. Gurney, D. V. Pynadath, V. Ustun, Spontaneous theory of mind for artificial intelligence, in: International Conference on Human-Computer Interaction, Springer, 2024, pp. 60–75.
[6] Y. Xu, J. Hu, Z. Zhao, Z. Duan, X. Sun, X. Yang, MultiAgentESC: a LLM-based multi-agent collaboration framework for emotional support conversation, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 4665–4681.
[7] P. Bellan, S. G. Haez, L. Sanna, S. Magnolini, M. Dragoni, Leveraging multi-agent systems for domain-pertinence query classification in informative chatbots, in: R. Bellazzi, J. M. Juarez Herrero, L. Sacchi, B. Zupan (Eds.), AIME, 2025, pp. 44–54.
[8] S. T. Völkel, Conversational agents with personality, Ph.D. thesis, LMU Munich, 2022.
[9] S. Roccas, Sagiv, et al., The big five personality factors and personal values, Personality and social psychology bulletin 28 (2002) 789–801.
[10] A. B. Kocaballi, S. Berkovsky, J. C. Quiroz, L. Laranjo, H. L. Tong, D. Rezazadegan, A. Briatore, E. Coiera, The personalization of conversational agents in health care: systematic review, Journal of medical Internet research 21 (2019).
[11] Y. Hwang, D. Shin, S. Baek, B. Suh, J. Lee, Applying the persona of user's family member and the doctor to the conversational agents for healthcare, 2021. arXiv:2109.01729.
[12] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, M. Matarić, Personality traits in large language models, 2023. arXiv:2307.00184.
[13] J. Liu, C. Symons, R. R. Vatsavai, Persona-based conversational AI: state of the art and challenges, 2022. arXiv:2212.03699.
[14] T. Hu, N. Collier, Quantifying the persona effect in LLM simulations, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 10289–10307.
[15] A. Liu, M. Diab, D. Fried, Evaluating large language model biases in persona-steered generation, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 9832–9850.
[16] T. Ge, X. Chan, X. Wang, D. Yu, H. Mi, D. Yu, Scaling synthetic data creation with 1,000,000,000 personas, arXiv (2024).
[17] L. Newsham, D. Prince, Personality-driven decision-making in LLM-based autonomous agents, in: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, 2025.
[18] P. Jandaghi, et al., Faithful persona-based conversational dataset generation with LLMs, in: Proceedings of the 6th Workshop on NLP4ConvAI 2024, 2024, pp. 114–139.
[19] D. X. Long, et al., Multi-expert prompting improves reliability, safety and usefulness of LLMs, in: Proceedings of the 2024 Conference on Empirical Methods in NLP, 2024.
[20] T. Liang, et al., Encouraging divergent thinking in LLMs through multi-agent debate, in: Proceedings of the 2024 Conference on EMNLP, 2024, pp. 17889–17904.
[21] W. Zhang, L. Zeng, Y. Xiao, Y. Li, C. Cui, Y. Zhao, R. Hu, Y. Liu, Y. Zhou, B. An, AgentOrchestra: a hierarchical multi-agent framework for general-purpose task solving, arXiv preprint arXiv:2506.12508 (2025).
[22] A. V. Singh, E. Rathbun, E. Graham, L. Oakley, S. Boboila, P. Chin, A. Oprea, Hierarchical multi-agent reinforcement learning for cyber network defense, in: AAMAS, Detroit, Michigan, USA, 2025. Extended Abstract.
[23] N. Nascimento, et al., Self-adaptive LLMs-based multiagent systems, in: 2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), IEEE, 2023, pp. 104–109.
[24] Z. Chai, et al., An expert is worth one token: Synergizing multiple expert LLMs as generalist via expert token routing, in: Proceedings of the 62nd Annual Meeting of the ACL (Volume 1: Long Papers), 2024.
[25] A. Greene, N. Blair, S. M. Aghabagher, S. Kumari, M. W. Schlund, A. Fedorov, V. D. Calhoun, X. Li, R. F. Silva, Ai psychiatrist assistant: an llm-based multi-agent system for depression assessment from clinical interviews, in: Machine Learning for Health 2025, 2025.
[26] D. Maas, M. Beigy, L. Genga, P. Van Gorp, Generating synthetic behavioral health data: A

multi-agent llm system (2026).
[27] D. Cabrera Lozoya, M. Conway, E. Sebastiano De Duro, S. D'Alfonso, Leveraging large language models for simulated psychotherapy client interactions: Development and usability study of client101, JMIR medical education 11 (2025) e68056.
[28] T. Scholich, M. Barr, S. Wiltsey Stirman, S. Raj, A comparison of responses from human therapists and large language model–based chatbots to assess therapeutic communication: Mixed methods study, JMIR Ment Health 12 (2025) e69709.
[29] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of mentalr iskes at iberlef 2024: Early detection of mental disorders risk in spanish, Procesamiento del lenguaje natural 73 (2024) 435–448.
[30] A. M. Mármol-Romero, P. Álvarez-Ojeda, A. Moreno-Mu, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ure, A. Montejo-Ráez, et al., Overview of mentalriskes at iberlef 2025: Early detection of addiction risk in spanish, Procesamiento del Lenguaje Natural 75 (2025) 425–440.
[31] A. Bonet-Jover, J. A. González-Barba, L. Chiruzzo, Overview of iberlef 2026: Natural language processing challenges for spanish and other iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.
[32] A. M. Mármol-Romero, F. Suárez-Maroto, R. García-Viedma, M. Hernández-López, F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, A. Montejo-Ráez, Overview of mentalriskes at iberlef 2026: Early detection of mental disorders risk in spanish, volume 77, 2026.
[33] P. Bellan, S. G. Haez, L. Sanna, S. Magnolini, M. Dragoni, Leveraging multi-agent systems for domain-pertinence query classification in informative chatbots, in: R. Bellazzi, J. M. Juarez Herrero, L. Sacchi, B. Zupan (Eds.), Artificial Intelligence in Medicine, 2025.
[34] R. L. Spitzer, K. Kroenke, J. B. W. Williams, B. Löwe, A brief measure for assessing generalized anxiety disorder: the GAD-7, Archives of Internal Medicine 166 (2006) 1092–1097.
[35] 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.
[36] D. T. Gillanders, H. Bolderston, F. W. Bond, M. Dempster, P. E. Flaxman, L. Campbell, S. Kerr, L. Tansey, K. Niven, C. Ferenbach, S. Masley, L. Roach, J. Lloyd, L. May, S. Clarke, B. Remington, The development and initial validation of the cognitive fusion questionnaire, Behavior Therapy 45 (2014) 83–101.
[37] M. Csikszentmihalyi, Beyond boredom and anxiety, Jossey-Bass behavioral science series, Jossey-Bass Publishers, San Francisco, CA, USA, 1975.
[38] P. Wason, J. Evans, Dual processes in reasoning?, Cognition 3 (1974) 141–154.
[39] J. Sweller, Cognitive load during problem solving: effects on learning, Cognitive Science 12 (1988) 257–285.
[40] S. Freud, A. Brill, The interpretation of dreams, Classics of World Literature, Wordsworth Editions, Ware, UK, 1997.
[41] D. A. Norman, S. W. Draper, User centered system design: new perspectives on human-computer interaction, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1986.
[42] A. Cooper, P. Saffo, The Inmates Are Running the Asylum, Sams Publishing, Indianapolis, IN, USA, 1999.
[43] A. Welivita, Y. Xie, P. Pu, A large-scale dataset for empathetic response generation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, 2021, pp. 1251–1264.
[44] M. Hosseini, C. Caragea, It takes two to empathize: One to seek and one to provide, in: Proceedings of the AAAI conference on artificial intelligence, 2021, pp. 13018–13026.
[45] H. Rashkin, E. M. Smith, M. Li, Y.-L. Boureau, Towards empathetic open-domain conversation models: a new benchmark and dataset, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence,

Italy, 2019, pp. 5370–5381.
[46] Y. Wu, Y. He, Y. Jia, R. Mihalcea, Y. Chen, N. Deng, Hi-ToM: a benchmark for evaluating higher-order theory of mind reasoning in large language models, in: Findings of the Association for Computational Linguistics: EMNLP, 2023, pp. 10691–10706.
[47] K. Shinoda, N. Hojo, K. Nishida, S. Mizuno, K. Suzuki, R. Masumura, H. Sugiyama, K. Saito, ToMATO: verbalizing the mental states of role-playing LLMs for benchmarking theory of mind, in: AAAI, 2025, pp. 1520–1528.
[48] H. Xu, R. Zhao, L. Zhu, J. Du, Y. He, OpenToM: a comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024, pp. 8593–8623.