AutoPsyC: Automatic Recognition of Psychodynamic Conflicts from Semi-structured Interviews with Large Language Models

Sayed Muddashir Hossain[1], Simon Ostermann[1], Patrick Gebhard[1], Cord Benecke[2], Josef van Genabith[1], Philipp Müller[1]

[1]DFKI, Saarbrücken, Germany
[2]University of Kassel, Kassel, Germany

{sayed_muddashir.hossain,simon.ostermann,patrick.gebhard}@dfki.de
{josef.van_genabith,philipp.mueller}@dfki.de

Abstract

Psychodynamic conflicts are persistent, often unconscious themes that shape a person's behaviour and experiences. Accurate diagnosis of psychodynamic conflicts is crucial for effective patient treatment and is commonly done via long, manually scored semi-structured interviews. Existing automated solutions for psychiatric diagnosis tend to focus on the recognition of broad disorder categories such as depression, and it is unclear to what extent psychodynamic conflicts which even the patient themselves may not have conscious access to could be automatically recognised from conversation. In this paper, we propose AutoPsyC, the first method for recognising the presence and significance of psychodynamic conflicts from full-length Operationalized Psychodynamic Diagnostics (OPD) interviews using Large Language Models (LLMs). Our approach combines recent advances in parameter-efficient fine-tuning and Retrieval-Augmented Generation (RAG) with a summarisation strategy to effectively process entire 90 minute long conversations. In evaluations on a dataset of 141 diagnostic interviews we show that AutoPsyC consistently outperforms all baselines and ablation conditions on the recognition of four highly relevant psychodynamic conflicts.

1 Introduction

Accurate and detailed analysis of clinical interviews is essential for effective psychodynamic diagnostics. In particular, Operationalized Psychodynamic Diagnostics (OPD) interviews (Force, 2008) serve as a cornerstone in psychodynamic assessment. A key aspect of OPD is the assessment of the patient's life-determining, often unconscious inner conflicts, such as conflicts relating to Dominance or Submissiveness, or to Self-value/esteem. Automated analysis of psychodynamic conflicts from clinical interviews has the potential to support clinicians, reduce manual work, enhance objectivity, and may even lay the groundwork for the delivery of diagnostic interviews by artificial agents. However, due to their long duration, low level of standardisation, and richness of information, semi-structured interviews pose unique challenges (Adams, 2010; Magaldi and Berler, 2020). Prior natural language processing (NLP) work has often focused on short interview excerpts and broad diagnostic categories (Low et al., 2020; Milintsevich et al., 2023). To the best of our knowledge, no previous work has addressed the recognition of fine-grained psychodynamic concepts from long semi-structured diagnostic interviews.

In this work, we introduce a novel approach for recognising the presence and significance of psychodynamic conflicts as classified in the OPD from full-length interviews. Our method combines recent advancements in parameter-efficient fine-tuning (Hu et al., 2022) and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) with a summarisation approach in order to process and classify long (90 min) semi-structured psychodynamic diagnostic interviews. In particular, we make use of a RAG framework to let the LLM access full-length interviews. To allow the model to effectively reason about the interview to be scored, we additionally prompt it with a summary of the interview. The classification is performed by an ensemble of four models, each of which was fine-tuned to analyse a specific temporal portion of an interview. To evaluate our approach, we make use of a dataset of 141 OPD interview recordings (Bock et al., 2016). Our approach consistently improves over baselines and ablation conditions. It is able to reach weighted F1 scores of 0.78 and 0.81 for the conflicts *Self-dependency and Dependency on Others*, and *Dominance or Submissiveness*. For the more challenging conflicts *Self-sufficiency* and *Self-value/esteem*, it is able to reach 0.59 and 0.58 F1, respectively.

Our specific contributions are threefold:

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 12–25

1. We present AutoPsyC[1], the first LLM-based method for the recognition of presence and severity of psychodynamic conflicts from full-length OPD interviews, thereby bridging the fields of psychodynamic diagnostics and advanced NLP.

2. We evaluate AutoPsyC on a dataset of 141 90 minute long OPD interviews, showing that AutoPsyC consistently outperforms all baselines and in-depth ablation comparisons.

3. We demonstrate that information contained in the middle sections of interviews is particularly informative for classifier training.

2 Related Work

2.1 Diagnostic Interviews in Psychotherapy and Psychiatry

Structured interviews, using standardized questions and scoring, improve psychiatric diagnosis reliability by reducing clinician bias. Tools like the Structured Clinical Interview for DSM-5 (SCID-5, First et al. (2016)) ensure DSM-aligned accuracy but require extensive training, while the Mini-International Neuropsychiatric Interview (MINI, Sheehan et al. (1998)) offers efficient screening at the cost of some diagnostic precision. The Structured Interview of Personality Organization (STIPO) is a structured interview designed to assess personality functioning based on Kernberg's object relations theory (Clarkin et al., 2007).

Unstructured interviews emphasize patient narratives and clinical intuition, enabling the exploration of unique experiences (Nordgaard et al., 2013). While fostering rapport and uncovering insights, their lack of standardization introduces variability and reduces reliability (Shea, 2016; Corbin and Morse, 2003; O'Brien and Tabaczynski, 2007; Widiger, 2008; Fava et al., 2024; Lenouvel et al., 2022). In this context, the PDM-2 provides a multidimensional diagnostic framework that emphasizes psychological functioning and personality organization over categorical symptom-based diagnosis (Lingiardi et al., 2015). Likewise, Malan's triangles offer a conceptual model for understanding intrapsychic conflict and resistance, rather than a formalized interview procedure (Malan, 1979).

Semi-structured interviews blend the structure of standardized questions with the flexibility to address emergent themes (Fava et al., 2024; Lenouvel et al., 2022; Adams, 2010; Brinkmann, 2014; Magaldi and Berler, 2020; Adeoye-Olatunde and Olenik, 2021). They have been shown to be particularly useful in complex cases like major depressive disorder (Dupuy et al., 2020). One example for a sem-structured format is the Core Conflictual Relationship Theme (CCRT) method to identify recurring interpersonal themes (Luborsky and Crits-Christoph, 1998). Operationalized Psychodynamic Diagnosis (OPD) uses semi-structured methods rooted in psychodynamic theory to assess self-experience, interpersonal relationships, and unconscious conflicts (Force, 2008; Cierpka et al., 2007). Unlike symptom-focused tools, OPD provides in-depth insights into personality organization and internal dynamics, aiding personalized therapeutic interventions. Research has demonstrated the clinical relevance of OPD within therapeutic settings (Cierpka et al., 2007; Benecke, 2024; Cierpka et al., 2001; Rudolf et al., 2004). Despite their importance, the automatic analysis of semi-structured interviews remains under-explored. In particular no previous work has attempted to automatically score OPD interviews.

2.2 Large Language Models for Psychiatric Diagnosis

The integration of NLP and machine learning in different aspects of mental health is a rapidly growing field of research (Le Glaz et al., 2021; Lindsay et al., 2021; Hossain et al., 2024). One particular focus of attention is the automated diagnosis of conditions such as depression or schizophrenia by analysing text, speech, and facial expressions (Barzilay et al., 2019; Low et al., 2020; Kishimoto et al., 2022; Oh et al., 2024; Milintsevich et al., 2023; Ettore et al., 2023). Tools like *Diagnostisches Expertensystem für psychische Störungen* (DIA-X-5) are being tested for consistency (Hoyer et al., 2020), while patient involvement is emphasized to ensure ethical use (Brederoo et al., 2021).

Recent work has shown that LLMs can be utilised to analyse complex human affect expression in conversation (Broekens et al., 2023; Müller et al., 2024), making them a promising candidate for applications in psychiatric disorders. Indeed, LLMs are increasingly applied in psychiatry, identifying linguistic markers of disorders from social media posts and clinical transcripts (Farruque et al., 2024; Lan et al., 2024; Zhang et al., 2024b). These models also assist in parsing unstructured EHR

[1]Code available at https://git.opendfki.de/philipp.mueller/autopsyc

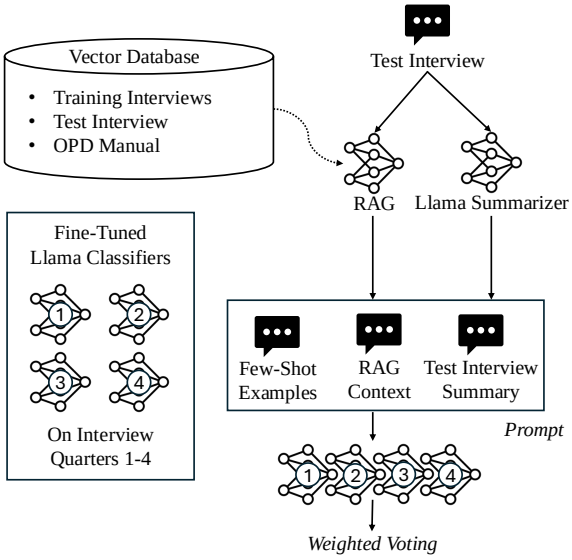

Figure 1: Overview over AutoPsyC.

notes for early diagnosis (Zhang et al., 2024b).

However, current automatic methods mainly detect broad disorder categories without capturing the more detailed and often unconscious factors explored in psychodynamic assessments. In particular, we are not aware of any approach to automatically recognise psychodynamic conflicts from clinical interviews.

2.3 Integrating Domain Knowledge in Large Language Models

The integration of domain-specific knowledge into LLMs, particularly through techniques like Retrieval-Augmented Generation (RAG), is transforming psychiatric applications by enabling models to access and apply current, relevant external information in real time (Lewis et al., 2020). Building on RAG, RAFT (Retrieval-Augmented Facilitation for Text) further optimizes domain-specific knowledge integration by prioritizing the most relevant medical literature and clinical guidelines (Zhang et al., 2024a). Other methods, such as parameter-efficient domain knowledge integration and the use of human-annotated features, also improve LLM performance in biomedical contexts (Ke et al., 2021; Kim et al., 2024). In our work, we utilise RAG techniques to present the first LLM-based system able to recognise psychodynamic conflicts expressed in full-length OPD interviews.

3 Method

A schematic overview of our method is shown in Figure 1. The classification of a test interview consists of three steps. In the first step, we employ LLaMA 3.1 (8B) (Grattafiori et al., 2024) to generate a summary of the interview. In the second step, we use this summary to build a prompt for the final classification, which is performed by an ensemble of LLaMa 3.1 (8B) models that are fine-tuned to different interview segments. The prompts for these specialised models contain 5 interview summaries from the training set, including ground truth (few-shot examples). Via a RAG framework (Lewis et al., 2020), each specialised model also has access to relevant sections of the OPD manual, as well as to the full test interview and all full interviews from the training set. In the third step, we combine the individual classifications obtained from the specialised models using a weighted voting scheme driven by a multinomial logistic regression.

3.1 Summarization Method

To obtain a focused representation of each interview, we first generate a summary using a LLaMA 3.1 8b model (Grattafiori et al., 2024). The summarization prompt includes an example summary excerpted from the OPD Manual, instructing the model to adhere to a consistent style that reflects the diagnostic criteria. In this way, the generated summaries capture the diagnostically relevant information while filtering out extraneous details.

3.2 Training Data Integration and RAG Setup

In addition to the test interview summary, the classification prompt also includes few-shot examples in the form of summaries of interviews from the training set with associated ground truth labels. We include one interview summary per ground truth class. To further ground the classification in a domain-specific context, we employ a Retrieval-Augmented Generation (RAG) framework. In particular, we upload the following information into the RAG vector database.

1. Training Interviews: We upload the full transcripts of all interviews from the training set without ground truth into the RAG knowledge base, pointing the model to them for retrieval. Adding ground truth information did not lead to improvements in preliminary experiments.

2. Test Interview: We also upload the full transcript of the test interview.

3. OPD Manual: The OPD-2 manual (Force, 2008) is organized into chapters correspond-

ing to its axes, providing detailed descriptions and examples of OPD tasks and classifications. For our purposes, we included excerpts from the chapter on conflicts (Axis III) and the introductory section where the axes are defined and explained. Preliminary tests showed that incorporating the entire OPD-2 manual into the retrieval-augmented generation (RAG) system did not improve model performance compared to using only the relevant excerpts.

This integration ensures that the classification model benefits from exemplars of each diagnostic class and explicit domain knowledge. The model is also able to access detailed information present in the full interview transcripts in case the summaries are inconclusive.

3.3 Classification Stage: Interview Segmentation and Finetuning

Given that OPD interviews are semi-structured—with diagnostic cues distributed throughout—we split each interview (or its summary) into k segments, where in our implementation $k = 4$ (each segment being roughly 5,000 words). In this way, each segments represents one quarter of the interview. For each segment, we fine-tune a separate Llama 3.1 (8B) model using LoRA (Hu et al., 2022), which allows for parameter-efficient adaptation. During training, we provide the model with a prompt including the segment summary, the RAG-augmented context (i.e., training interview summaries and manual excerpts), and an example for each of the five classes. The model is trained to output a probability distribution over the diagnostic classes. This process allows each fine-tuned model to capture the specific semantic and contextual nuances present in its corresponding interview segment.

3.4 Result Aggregation

After obtaining classification probabilities from each of the four fine-tuned models, we combine their outputs using a weighted voting scheme. Specifically, we train a multinomial logistic regression model that assigns a weight w_i to the prediction $p_i(c)$ of the i-th segment for class c. The final predicted class $\hat{y}$ is computed as:

$$\hat{y} = \arg\max_{c} \sum_{i=1}^{k} w_i \, p_i(c) \tag{1}$$

where $k = 4$ in our implementation.

4 Dataset

The Kassel dataset (Bock et al., 2016), utilized in this study, comprises 141 participants recorded during Operationalized Psychodynamic Diagnostics (OPD) interviews.

4.1 Participants

The dataset includes both male (n = 21) and female (n = 120) participants, aged between 18 and 57 years. Among them, 64 were inpatients diagnosed with at least one DSM-IV (Association, 2000) disorder, while 20 were healthy controls. The remaining participants had diverse diagnostic categories, including somatoform disorders (n = 22), borderline personality disorder (n = 19), depression (n = 18), and eating disorders such as anorexia (n = 14) and bulimia (n = 14). Anxiety disorders were observed in 13 participants. The inclusion criteria required informed consent, age above 18 years, and the absence of acute psychosis or schizophrenia.

4.2 Data Collection

Each participant underwent a clinical interview based on the framework of Operationalized Psychodynamic Diagnostics (Force, 2008). The interviews were carried out by a team of two male and two female interviewers (Bock et al., 2016), all of whom were certified and trained in OPD application. The sessions, with an average duration of approximately 90 minutes, were recorded using split-screen technology to capture both the participants and interviewers. Both the interviewer and the interviewee were equipped with microphones to ensure clear audio capture. The audio from each session was transcribed verbatim into text by research staff or trained transcribers (Bock et al., 2016; Vierl et al., 2023). These recordings provided the foundational data for subsequent analyses of behavioural and contextual elements.

4.3 Clinical Ground Truth

The dataset includes scores for Axes I-V of the OPD system. In the scope of this paper, we focus on Axis III, which captures the patient's life-determining (un)conscious inner conflicts. In

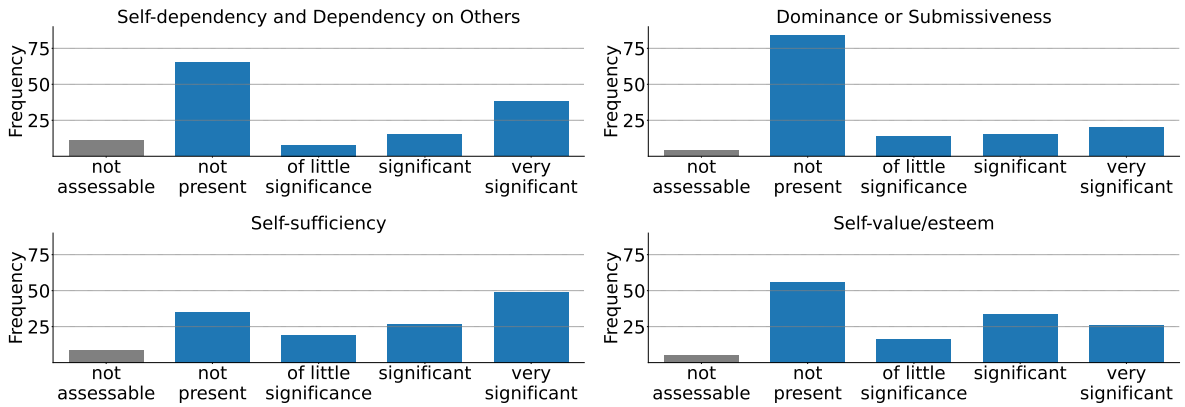

Figure 2: Class distributions for the psychodynamic conflicts investigated in this study.

particular, these are: *Conflicts related to self-dependency and dependency on others*, *Conflicts associated with dominance or submissiveness*, *Conflicts revolving around self-sufficiency*, *Self-value and self-esteem conflicts*, *Oedipal conflicts*, *Identity-related conflicts*. Each of these conflicts is rated with a five-class classification scheme. The classes are *Not assessable, Not present, Of little significance, Significant, Very significant*. Their detailed description can be found in A.3.

We decided to omit the conflicts *Oedipal Conflict* and *Conflicts Related to Identity* from further analysis, as these conflicts were diagnosed in only a few instances, making a robust evaluation of predictions infeasible. For instance, in the case of *Oedipal Conflict*, 120 out of 141 instances were labeled as *not present* (see Appendix for further details). Figure 2 illustrates the class distribution of the remaining conflicts. While *not present* is the most prevalent class for every conflict except Self-sufficiency, in all cases a significant portion of participants exist for whom the respective conflict is at least present with little significance. In the following we provide a concise explanation of the four conflicts included in our analysis.

Conflicts related to self-dependency and dependency on others refer to the tension between striving for autonomy and seeking support from others, often leading to struggles between independence and fear of isolation. *Conflicts associated with dominance or submissiveness* involve power dynamics in interpersonal relationships, where individuals may oscillate between asserting control and yielding to authority, potentially resulting in power struggles or passivity. *Conflicts revolving around self-sufficiency* pertain to the balance between the need for care and the desire for independence, with individuals experiencing inner turmoil when their reliance on others contradicts their self-reliance. *Self-value and self-esteem conflicts* center on an individual's sense of worth, encompassing struggles with feelings of inadequacy or inferiority, often manifesting in compensatory behaviours aimed at reinforcing self-image. The conflicts were classified based on the overall interview. When binarizing the conflicts by treating *not assessable* and *not present* as "No," and all other categories as "Yes," only three cases did not present with any of the conflicts.

5 Experiments

In this section, we describe the evaluation protocol and baselines.

5.1 Evaluation Protocol

The dataset was partitioned into five fixed folds using stratified 5-fold cross-validation to maintain the proportional representation of key demographic and diagnostic variables. This was achieved by utilizing the `StratifiedKFold` module from the `scikit-learn` library (Pedregosa et al., 2011), with stratification based on *Gender* and *Diagnosis*. The stratification process guarantees a fair distribution of these attributes across all folds. This consistency was preserved across all experiments involving the 5-fold cross-validation framework.

To evaluate all models and baselines, we make use of the weighted F1 score. The weighted F1 score accounts for class imbalance by weighting classes proportionally to their prevalence, ensuring robust evaluation of both frequent and rare diagnostic categories. It balances precision (avoiding over-

	Self-dep. & others-dep.	Dom. or sub.	Self-suff.	Self-val. & self-est.
Naive Baselines				
Demographic	0.31 (±0.01)	0.46 (±0.02)	0.31 (±0.03)	0.26 (±0.01)
Random	0.30 (±0.00)	0.33 (±0.00)	0.20 (±0.00)	0.23 (±0.00)
No Training Data in VDB, No Fine-tuning				
w/o Manual	0.51 (±0.01)	0.63 (±0.01)	0.39 (±0.02)	0.43 (±0.03)
w/o Test Interv. in VDB	0.46 (±0.01)	0.60 (±0.01)	0.46 (±0.02)	0.48 (±0.01)
w/o Test Interv. Summary	0.53 (±0.01)	0.64 (±0.01)	0.42 (±0.02)	0.46 (±0.01)
w/o Few-shot Examples	0.65 (±0.01)	0.68 (±0.01)	0.53 (±0.03)	0.48 (±0.02)
AutoPsyC	0.68 (±0.01)	0.70 (±0.02)	0.55 (±0.01)	0.48 (±0.02)
Training Data in VDB (Unlabelled), No Fine-tuning				
w/o Manual	0.48 (±0.03)	0.61 (±0.02)	0.43 (±0.02)	0.49 (±0.02)
w/o Test Interv. in VDB	0.50 (±0.01)	0.60 (±0.01)	0.45 (±0.01)	0.50 (±0.01)
w/o Test Interv. Summary	0.62 (±0.01)	0.69 (±0.02)	0.47 (±0.01)	0.52 (±0.02)
w/o Few-shot Examples	0.68 (±0.02)	0.73 (±0.01)	0.57 (±0.01)	0.50 (±0.04)
AutoPsyC	0.70 (±0.01)	0.74 (±0.02)	0.58 (±0.01)	0.50 (±0.02)
Training Data in VDB (Unlabelled), Fine-tuning				
w/o Test Interv. Summary & Manual & Train Interv. in VDB	0.65 (±0.04)	0.68 (±0.01)	0.49 (±0.02)	0.47 (±0.02)
w/o Test Interv. Summary & Manual	0.69 (±0.02)	0.72 (±0.03)	0.50 (±0.01)	0.49 (±0.01)
w/o Test Interv. Summary & Weighted Voting	0.73 (±0.01)	0.75 (±0.02)	0.56 (±0.01)	0.53 (±0.02)
w/o Test Interv. Summary & Ensemble	0.72 (±0.02)	0.74 (±0.01)	0.55 (±0.02)	0.52 (±0.01)
w/o Manual & Train Interv. in VDB	0.68 (±0.02)	0.72 (±0.01)	0.51 (±0.02)	0.49 (±0.01)
w/o Weighted Voting	0.75 (±0.02)	0.78 (±0.01)	0.55 (±0.02)	0.57 (±0.01)
w/o Ensemble	0.71 (±0.02)	0.74 (±0.01)	0.56 (±0.02)	0.55 (±0.01)
w/o Train Interv. in VDB	0.74 (±0.02)	0.77 (±0.01)	0.56 (±0.02)	0.55 (±0.01)
w/o Manual	0.72 (±0.01)	0.74 (±0.02)	0.54 (±0.01)	0.52 (±0.02)
w/o Test Interv. in VDB	0.70 (±0.02)	0.73 (±0.01)	0.53 (±0.02)	0.50 (±0.01)
w/o Test Interv. Summary	0.75 (±0.01)	0.80 (±0.02)	0.57 (±0.01)	0.57 (±0.02)
w/o Few-shot Examples	0.73 (±0.01)	0.74 (±0.02)	0.55 (±0.01)	0.51 (±0.02)
AutoPsyC	**0.78 (±0.02)**	**0.81 (±0.01)**	**0.59 (±0.02)**	**0.58 (±0.01)**

Table 1: Average Weighted F1-Scores with 95% Confidence Intervals.

pathologizing) and recall (preventing missed conflicts), aligning with clinical priorities. To robustly estimate performance, we repeated all experiments several times and report the average weighted F1 score across all runs. In the case of experiments involving LLMs, we average across 100 runs, and in the case of the computationally less expensive baseline experiments, we average across 1000 runs. In addition to the averages, we also report their 95% confidence interval.

5.2 Baselines

We implement two simple baselines: a *Demographic Baseline* and a *Random Baseline*.

Demographic Baseline. This baseline employs a neural network classifier using demographic attributes such as gender, clincial diagnosis group, and binned age as input features. Numerical features were normalized by subtracting the mean and dividing by the standard deviation, while categorical features were converted into numerical representations. The neural network, implemented in PyTorch (Paszke et al., 2019), consists of three fully connected layers with ReLU activations. It was trained for 30 epochs using cross-entropy loss and the Adam optimizer.

Random Baseline. The random baseline leverages the `DummyClassifier` module from `scikit-learn`, configured with the `stratified` strategy. This classifier generates predictions by randomly assigning labels based on the class distribution of the training set. This random classifier serves as a naive baseline, highlighting the minimum expected performance for the classification task.

6 Results and Discussion

6.1 Overview

Table 1 summarises our weighted F_1 scores across the four psychodynamic conflicts. To more easily navigate the table, we partition the different ablation conditions into three cases, based on whether unlabelled training data is encorporated in the RAG framework and based on whether fine-tuning is performed with labelled training data. Ablations are always named relative to the partition their are in. For example, the ablation *w/o Manual* in the partition *No Training Data in VDB, No Fine-tuning*

describes an ablation condition without training data integration into the RAG framework, without fine-tuning, and without integration of the OPD Manual in the RAG framework.

We observe that our full method (AutoPsyC), which combines our summarisation strategy with weighted voting across fine-tuned, temporally specialised models, as well as domain knowledge integration into the RAG framework, consistently outperforms all baselines and ablation conditions. As illustrated in Figure 3, models fine-tuned on the middle segments of the interviews consistently outperform those focusing on earlier or later sections. Moreover, Figure 4 indicates that deviating from four total models or partitioning the interviews into fewer or more than four segments leads to a noticeable drop in overall performance. Finally, we present an analysis of gender fairness of our model. Overall low values of Conditional Demographic Disparity (CDD) indicate no major gender-related biases (Table 2).

6.2 Which Model Configuration Works Best?

Our experimental results demonstrate the effectiveness of combining our summarisation strategy with Retrieval-Augmented Generation (RAG), instruction tuning and section-wise model specialization for psychodynamic conflict classification in clinical interviews. As can be seen in (Table 1), AutoPsyC achieves superior performance across all four conflict categories, with weighted F1-scores ranging from 0.58 to 0.81. This represents a substantial improvement over both naive baselines (Demographic: 0.26–0.46; Dummy: 0.20–0.33) and non-instruction-tuned variants (0.50–0.74).

Our detailed ablation experiments indicates that AutoPsyC effectively integrates all available information. We can observe a large decrease in performance when the test interview transcript is removed from the vector database (0.50-0.73 F1). This indicates that our model indeed makes use of the full test interview transcript that is provided via the RAG framework to fill in information missing in the interview summary. At the same time, we see that it does profit from the test interview summary, with losses of up to 0.04 F1 when the summary is removed. We furthermore observe a clear loss in performance when the OPD manual is removed from the vector data base (0.52-0.74 F1), and a slightly lower loss in performance when the training set interviews are removed from the database (0.55-0.77 F1). This indicates that even when using fine-tuned classification model, domain knowledge integration via the RAG setup is still helpful. The weighted voting mechanism using multinomial logistic regression provides moderate but consistent performance gains (0.58–0.81 F1 vs. 0.55–0.78 for unweighted aggregation), suggesting that different interview sections contribute asymmetrically to conflict identification.

One general observation we can make is that fine-tuning leads to greater robustness w.r.t. other ablation conditions. E.g. removing the OPD Manual from the vector database leads only to a moderate loss in performance when fine-tuned classification models are used (0.58-0.81 F1 vs. 0.52-0.74 F1). In contrast, for the case of no fine-tuning, the losses are more dramatic (0.50-0.74 F1 vs. 0.43-0.61 F1).

6.3 Which Interview Section is most useful?

To further investigate which sections of the interviews are most informative fine-tuning classification models, we investigate the performance of our four pretrained models singled out across all conflicts (see Figure 3). The results indicate that the models fine-tuned using the middle sections of the interviews outperform those tuned with other sections. After a careful examination of the interviews, we observed that, in the quarter 2 & 3, the interviewees often share information more closely related to their condition and situation. An excerpt of the interview can be found in Appendix A.1.

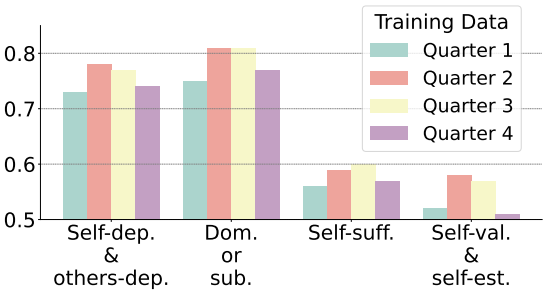

Figure 3: Performance of the four models across all classes.

6.4 Additional Experiments: How Fair is the Model?

Fairness concerning gender is a critical issue in psychiatric diagnoses when using machine learning algorithms, as biases in training data or model predictions can lead to systematic disparities in diagnostic outcomes. For instance, a study by Mosteiro et al.

	Self-dep. & others-dep.	*Dom. or sub.*	*Self-suff.*	*Self-val. & self-est.*
not assessable	0.0031	0.0008	0.0053	0.0044
not present	0.0042	0.0023	0.0034	0.0014
of little significance	0.0018	0.0011	0.0019	0.0020
significant	0.0025	0.0021	0.0054	0.0009
very significant	0.0029	0.0015	0.0010	0.0027

Table 2: One-vs.-rest CDD values for each class across four conflicts.

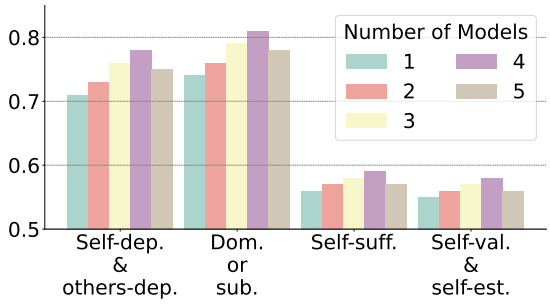

Figure 4: Performance of Different Number of Models

(2022) found that gender played an unexpected role in predictions related to benzodiazepine administration, potentially biasing the model's decisions.

We evaluate fairness with respect to the gender attribute using Conditional Demographic Disparity (CDD) (Wachter et al., 2021). CDD quantifies the difference in expected outcomes across demographic groups, with values closer to zero indicating fairer conditions.

$$\text{CDD} = \mathbb{E}[\hat{y} \mid \text{male}, y] - \mathbb{E}[\hat{y} \mid \text{female}, y] \quad (2)$$

Since we have more than two classes we compute CDD in one-vs.-rest fashion. We compute

$$\text{CDD}_k = \mathbb{E}\big[\hat{p}^k \mid \text{male},\, y\big] \;-\; \mathbb{E}\big[\hat{p}^k \mid \text{female},\, y\big] \quad (3)$$

for class k. This formulation reduces the multiclass problem to a "one vs. rest" scenario by focusing on a single predicted probability $\hat{p}^k$. If class k is deemed the "positive" class, the binary-based fairness thresholds from (Wachter et al., 2020) can be applied to CDD_k.

Overall, the CDD values reported in Table 2 are relatively low (all below 0.006), as indicated by prior studies (Wachter et al., 2021; Koumeri et al., 2023; Wachter et al., 2020), suggesting minimal gender-based disparity across the different classes. Wachter et al. (2021) suggest that CDD values below 0.01 indicate minimal demographic disparity, while Koumeri et al. (2023) and Wachter et al. (2020) provide empirical evidence that values above 0.02-0.05 often indicate fairness concerns. It is important to note that this fairness evaluation is not able to account for potential biases that are already present in the ground truth annotations on the dataset.

6.5 Ethical Considerations and Impact

The automation of psychodynamic diagnostics using NLP and machine learning presents both opportunities and ethical challenges. While enhancing objectivity, efficiency, and accessibility, its implementation requires careful ethical scrutiny to ensure responsible use in mental health care. Psychodynamic interviews contain sensitive data, necessitating strong anonymization and compliance with privacy regulations such as GDPR and HIPAA. Additionally, automated diagnostics may reflect biases present in training data, leading to disparities across demographic groups. Continuous bias auditing and fairness assessments are essential to mitigate these risks and ensure equitable model performance.

Automated tools should complement, not replace, human expertise. AutoPsyC could serve as a supplementary tool for therapists during psychodynamic interviews, acting as a "second-eye" to enhance clinical decision-making (American Psychological Association, 2025). Additionally, AutoPsyC could be integrated into social interactive agents, chat applications, and telepsychiatry platforms, providing complementary therapeutic support (Smith et al., 2019). Furthermore, it could be utilized in psychological training tools to enhance the proficiency of conducting psychodynamic interviews (American Psychological Association, 2023).

Psychodynamic diagnostics involve complex interpretations that extend beyond text-based pattern recognition. Thus, model outputs must be interpretable, allowing clinicians to integrate them into their assessments. Future research should prioritize explainability and transparency in AI-driven diagnostics. As AI applications in mental health

expand, concerns arise regarding consent, misuse, and potential stigmatization in non-clinical settings. Interdisciplinary collaboration among clinicians, ethicists, and policymakers is needed to safeguard patient autonomy and well-being.

7 Conclusion

We present a novel framework for automated conflict classification in psychodynamic interviews, achieving clinically relevant performance through three key innovations: (1) domain-adapted instruction tuning using segmented interview data, (2) RAG-enhanced contextual understanding through OPD Manual and other interview integration, and (3) confidence-weighted aggregation of specialized section models.

These results suggest that LLMs can be effectively adapted for complex psychiatric coding tasks when combined with domain-specific knowledge retrieval and structured interview analysis. The demonstrated technique for identifying diagnostically salient interview segments (quarters 2 & 3) offers methodological insights for computational psychiatry research. Future work should explore applications to other OPD Axes and integration with multimodal clinical data.

8 Limitations

While promising, our approach has several limitations. First, the dataset size (n=141 interviews) may limit generalizability, particularly for rare conflict subtypes. Second, the complex pipeline (RAG, summarization, 4 specialized models) incurs significant computational costs compared to monolithic models. Third, performance variation across conflict categories (0.58–0.81 F1) suggests task-specific optimization needs, particularly for *Self-sufficiency* classification.

The reliance on manual OPD Manual examples for summarization introduces potential annotation bias, and the gender fairness analysis does not account for non-binary identities. Additionally, our stratified sampling based on diagnosis and gender may not fully capture all confounding demographic factors. We focus on a single summarization approach, as our primary goal is to establish a proof of concept for automated OPD scoring. While alternative summarization methods could be explored in future work, this choice allows us to maintain methodological consistency and provide a clear baseline for comparison.

We split the interview into four parts based on word counts, which does not fully account for the semi-structured nature of our interviews. Future work could focus on automatically detecting interview segments for the fine-tuning process.

Our fairness analysis showed minimal gender-based disparity in predictions. There are however many other ways in which our model may be biased. On the interview dataset we utilised, we did not have access to e.g. information on socioeconomic status or education. Further variations are such as cultural background are not sufficiently covered by the dataset as it was recorded with German-speaking people in Europe. This geographic and cultural constraint represents another key limitation of our study. It remains unclear, whether our approach would also work in very different cultural contexts.

Future research should address these limitations through larger multicentre datasets (König et al., 2022), lightweight model architectures, and explicit modeling of clinician raters' variance. The current implementation also requires further validation for real-time clinical deployment, including robustness testing against speech recognition errors and patient dialect variations. A more detailed investigation of how AutoPsyC handles defensive processes (Freud, 1936) remains an area for future research.

References

Eike Adams. 2010. The joys and challenges of semi-structured interviewing. *Community Practitioner*, 83(7):18–22.

Omolola A Adeoye-Olatunde and Nicole L Olenik. 2021. Research and scholarly methods: Semi-structured interviews. *Journal of the american college of clinical pharmacy*, 4(10):1358–1367.

American Psychological Association. 2023. Ai is changing every aspect of psychology. here's what to watch for. *Monitor on Psychology*.

American Psychological Association. 2025. Artificial intelligence in mental health care. *American Psychological Association*.

American Psychiatric Association. 2000. *Diagnostic and Statistical Manual of Mental Disorders (4th ed., Text Revision)*. American Psychiatric Association, Washington, DC.

Ran Barzilay, Nadav Israel, Amir Krivoy, Roi Sagy, Shiri Kamhi-Nesher, Oren Loebstein, Lior Wolf, and Gal Shoval. 2019. Predicting affect classification

in mental status examination using machine learning face action recognition system: a pilot study in schizophrenia patients. *Frontiers in Psychiatry*, 10:446117.

Cord Benecke. 2024. 30 jahre operationalisierte psychodynamische diagnostik – neuerungen in der opd-3. *Psychotherapeutenjournal*, 23(1):36–46.

Astrid Bock, Eva Huber, and Cord Benecke. 2016. Levels of structural integration and facial expressions of negative emotions. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, 62:224–238.

SG Brederoo, FG Nadema, FG Goedhart, AE Voppel, JN De Boer, J Wouts, S Koops, and IEC Sommer. 2021. Implementation of automatic speech analysis for early detection of psychiatric symptoms: what do patients want? *Journal of psychiatric research*, 142:299–301.

Svend Brinkmann. 2014. Unstructured and semi-structured interviewing. *The Oxford handbook of qualitative research*, 2:277–299.

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th international conference on affective computing and intelligent interaction (ACII)*, pages 1–8. IEEE.

M. Cierpka, C. Grande, J. Rudolf, B. Stasch, and M. von der Tann. 2001. The operationalized psychodynamic diagnostics system: Clinical relevance, reliability, and validity. *Psychopathology*, 34(4):209–220.

Manfred Cierpka, Tilman Grande, Gerd Rudolf, M Von Der Tann, and Michael Stasch. 2007. The operationalized psychodynamic diagnostics system: clinical relevance, reliability and validity. *Psychopathology*, 40(4):209–220.

John F. Clarkin, Eve Caligor, Barry L. Stern, and Otto F. Kernberg. 2007. The structured interview of personality organization (stipo): A preliminary report. *Journal of Personality Assessment*, 88(1):69–83.

Juliet Corbin and Janice M Morse. 2003. The unstructured interactive interview: Issues of reciprocity and risks when dealing with sensitive topics. *Qualitative inquiry*, 9(3):335–354.

Lucile Dupuy, Jean-Arthur Micoulaud-Franchi, Hélène Cassoudesalle, Orlane Ballot, Patrick Dehail, Bruno Aouizerate, Emmanuel Cuny, Etienne de Sevin, and Pierre Philip. 2020. Evaluation of a virtual agent to train medical students conducting psychiatric interviews for diagnosing major depressive disorders. *Journal of Affective Disorders*, 263:1–8.

Eric Ettore, Philipp Müller, Jonas Hinze, Matthias Riemenschneider, Michel Benoit, Bruno Giordana, Danilo Postin, Rene Hurlemann, Amandine Lecomte, Michel Musiol, et al. 2023. Digital phenotyping for differential diagnosis of major depressive episode: narrative review. *JMIR mental health*, 10:e37225.

Nawshad Farruque, Randy Goebel, Sudhakar Sivapalan, and Osmar R Zaïane. 2024. Depression symptoms modelling from social media text: an llm driven semi-supervised learning approach. *Language Resources and Evaluation*, pages 1–29.

Giovanni A Fava, Nicoletta Sonino, David C Aron, Richard Balon, Carmen Berrocal Montiel, Jianxin Cao, John Concato, Ajandek Eory, Ralph I Horwitz, Chiara Rafanelli, et al. 2024. Clinical interviewing: an essential but neglected method of medicine. *Psychotherapy and psychosomatics*, 93(2):94–99.

Michael B. First, Janet B. W. Williams, Rhonda S. Karg, and Robert L. Spitzer. 2016. Structured clinical interview for dsm-5® disorders—clinician version (scid-5-cv).

OPD Task Force. 2008. *Operationalized psychodynamic diagnosis OPD-2: Manual of diagnosis and treatment planning*. Hogrefe Publishing GmbH.

Anna Freud. 1936. Das ich und die abwehrmechanismen.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sayed Muddashir Hossain, Jan Alexandersson, and Philipp Müller. 2024. M3TCM: Multi-modal multi-task context model for utterance classification in motivational interviews. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

Jana Hoyer, Catharina Voss, Jens Strehle, John Venz, Lars Pieper, Hans-Ulrich Wittchen, Stefan Ehrlich, and Katja Beesdo-Baum. 2020. Test-retest reliability of the computer-assisted dia-x-5 interview for mental disorders. *BMC psychiatry*, 20:1–16.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Pengfei Ke, Xiaoman Ji, Weizhi Wang, and Min Sun. 2021. Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3876–3887.

Hyunji Kim, Byeongchang Choi, Hyeongu Cho, Junwon Park, Eunji Kim, Sungwon Kim, and Jiyeon Kang. 2024. Towards understanding counseling conversations: Domain knowledge and large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1675–1687.

Taishiro Kishimoto, Hironobu Nakamura, Yoshinobu Kano, Yoko Eguchi, Momoko Kitazawa, Kuo-ching Liang, Koki Kudo, Ayako Sento, Akihiro Takamiya, Toshiro Horigome, et al. 2022. Understanding psychiatric illness through natural language processing (underpin): Rationale, design, and methodology. *Frontiers in Psychiatry*, 13:954703.

Alexandra König, Philipp Müller, Johannes Tröger, Hali Lindsay, Jan Alexandersson, Jonas Hinze, Matthias Riemenschneider, Danilo Postin, Eric Ettore, Amandine Lecomte, et al. 2022. Multimodal phenotyping of psychiatric disorders from social interaction: Protocol of a clinical multicenter prospective study. *Personalized Medicine in Psychiatry*, 33:100094.

L.K. Koumeri, M. Legast, Y. Yousefi, and K. Vanhoof. 2023. Compatibility of fairness metrics with EU non-discrimination laws: Demographic parity & conditional demographic disparity. *arXiv preprint*.

Xiaochong Lan, Yiming Cheng, Li Sheng, Chen Gao, and Yong Li. 2024. Depression detection on social media with large language models. *arXiv preprint arXiv:2403.10750*.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, 23(5):e15708.

Eric Lenouvel, Camelia Chivu, Janet Mattson, John Q Young, Stefan Klöppel, and Severin Pinilla. 2022. Instructional design strategies for teaching the mental status examination and psychiatric interview: a scoping review. *Academic Psychiatry*, 46(6):750–758.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Hali Lindsay, Philipp Mueller, Nicklas Linz, Radia Zeghari, Mario Magued Mina, Alexandra König, and Johannes Tröger. 2021. Dissociating semantic and phonemic search strategies in the phonemic verbal fluency task in early dementia. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 32–44.

Vittorio Lingiardi, Nancy McWilliams, Robert F. Bornstein, Francesco Gazzillo, and Robert M. Gordon. 2015. The psychodynamic diagnostic manual version 2 (pdm-2): Assessing patients for improved clinical practice and research. *Psychoanalytic Psychology*, 32(1):94–115.

Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope investigative otolaryngology*, 5(1):96–116.

Lester Luborsky and Paul Crits-Christoph. 1998. *Who Will Benefit from Psychotherapy?: Predicting Therapeutic Outcomes*. Basic Books, New York, NY.

Danielle Magaldi and Matthew Berler. 2020. Semi-structured interviews. *Encyclopedia of personality and individual differences*, pages 4825–4830.

David H. Malan. 1979. *Individual Psychotherapy and the Science of Psychodynamics*. Butterworth-Heinemann, Oxford, UK.

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2023. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1):4.

Pablo Mosteiro, Jesse Kuiper, Judith Masthoff, Floortje Scheepers, and Marco Spruit. 2022. Bias discovery in machine learning models for mental health. *Information*, 13(5):237.

Philipp Müller, Alexander Heimerl, Sayed Muddashir Hossain, Lea Siegel, Jan Alexandersson, Patrick Gebhard, Elisabeth André, and Tanja Schneeberger. 2024. Recognizing emotion regulation strategies from human behavior with large language models. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*.

Julie Nordgaard, Louis A. Sass, and Josef Parnas. 2013. The psychiatric interview: validity, structure, and subjectivity. *Eur Arch Psychiatry Clin Neurosci*, 263:353–64.

Jihoon Oh, Taekgyu Lee, Eun Su Chung, Hyonsoo Kim, Kyongchul Cho, Hyunkyu Kim, Jihye Choi, Hyeon-Hee Sim, Jongseo Lee, In Young Choi, et al. 2024. Development of depression detection algorithm using text scripts of routine psychiatric interview. *Frontiers in psychiatry*, 14:1256571.

William H O'Brien and Tracy Tabaczynski. 2007. Unstructured interviewing. *Handbook of clinical interviewing with children*, pages 16–29.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: an imperative style.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

J. Rudolf, C. Grande, and M. Cierpka. 2004. Operationalized psychodynamic diagnosis in planning and conducting psychotherapy. *Psychotherapy Research*, 14(3):295–308.

Shawn Christopher Shea. 2016. *Psychiatric interviewing: The art of understanding: A practical guide for psychiatrists, psychologists, counselors, social workers, nurses, and other mental health professionals*. Elsevier Health Sciences.

David V Sheehan, Yves Lecrubier, K Harnett Sheehan, Patricia Amorim, Juris Janavs, Emmanuelle Weiller, Thierry Hergueta, Roxy Baker, Geoffrey C Dunbar, et al. 1998. The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *Journal of clinical psychiatry*, 59(20):22–33.

Anthony C. Smith, Penny M. N. Probst, Len Gray, Richard G. C. Johnston, Peter C. Wootton, and David G. C. Williams. 2019. Role of artificial intelligence within the telehealth domain. *Yearbook of Medical Informatics*, 28(1):162–167.

Larissa Vierl, Charlotte Von Bremen, York Hagmayer, Cord Benecke, and Christian Sell. 2023. How are psychodynamic conflicts associated with personality functioning? a network analysis. *Frontiers in Psychology*, 14:1152150.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review*, 123:765–810.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567.

Thomas A Widiger. 2008. Clinical interviews. *Evidence-based outcome research: A practical guide to conducting randomized controlled trials for psychosocial interventions*, pages 47–65.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024a. Raft: Adapting language model to domain specific rag.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024b. When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. *arXiv preprint arXiv:2402.13276*.

A Appendix

A.1 Excerpt from the Interview Transcript

Beginning

Interviewer: So, let us begin with the second part. I will ask you various questions from different areas.

Interviewee: Hmm.

Interviewer: About your present life, your past, relationships, work life, etc. Yes. I would like to start by asking you to describe what is currently the most burdensome for you.

Interviewee: At the moment?

Interviewer: Yes, it can be anything.

Interviewee: (Exhales) Ahm. (-) In general, I am actually doing quite well. However, I must say that things that have burdened me in the past, especially over the past three years, have now become less significant. What currently affects me the most is the situation at home. My parents are about to get divorced, and that has been the most difficult thing in my life so far. I must say, it has also been very stressful for me, but I am slowly managing it quite well.

Middle

Interviewee: The period was simply long. I was 20 when it all started, and I would say that only in the last few months have I truly felt lighter inside. For about a year, things have been steadily improving, but before that, I felt terrible. At home, it was a crisis. My mother was struggling—she barely ate, she was just existing. That made me very sad because I am someone who tries to keep everyone together. Given my age, I was able to grasp everything more clearly. I always spoke with everyone, tried to mediate, and made sure we somehow lived through it. But it was

simply too much. (Claps hands on the table.)

I do not regret anything, or at least not much, except for the lingering aftereffects, which sometimes scare me. But otherwise, I would do it all over again. It just went too far. There were long periods where I barely met anyone or made any plans. If someone invited me out, I would always say no because I had to check on my mother to see if she was alone.

It was a responsibility that suddenly fell upon me. I would not say that it was forced upon me—I took it on willingly. That is simply the kind of person I am. If I see someone struggling, I cannot ignore it. I am very attached to my family.

There were times, for example, at Easter, when my mother just drove off. I could see in her eyes that she did not want to live anymore. She says the same thing even now. Back then, it was even stronger—she simply did not want to go on. She just got into the car and drove away. (Shocked and saddened.) It was simply terrible.

At first, I wanted to prevent the separation, of course. As a child, you never want your parents to separate. But later, it was just about minimizing the damage. I lost count of how many times I sat there listening, trying to mediate. I took on the role of always being there. But at some point, it was just too much.

I still managed to get through it, though sometimes I look back and wonder how I did it. I held up well, except for my university studies, where I had some setbacks. That was where the burden really showed. The emotional toll and the time commitment were simply too much.

End

Interviewee: I actually feel much better now. I have accepted everything as it is. I am a realistic person. I do not try to convince myself of things that do not exist. I walk through life with open eyes. I see what is happening around me. I know the divorce statistics.

Interviewer: But until recently, they did not matter to you.

Interviewee: No.

Interviewer: (Laughs.)

Interviewee: (Laughs as well.) Yes, because within those four walls, everything was fine. That was my foundation, my roots, where I came from. It was intact.

Interviewer: But now that has changed.

Interviewee: Yes. And I know that no matter how well things may seem to be going, there is always the risk that it could fall apart. That belief, that certainty I once had, is gone. I used to truly believe in lasting relationships. But now, if you ask me whether I think a relationship will last a lifetime, I no longer believe that. It is a sad realization.

Interviewer: It sounds as if a vision or a dream has been lost.

Interviewee: Yes, definitely. No doubt about it.

A.2 Example Prompt for *Self-dependency and Dependency on Others*

Context: Relationships and attachments are of existential importance in every person's life. They span the opposing poles of striving for close relationships and symbiotic proximity (dependency) and striving for well-developed independence and clear distance (powerful individuation). Individuation and dependency are fundamental elements

of human life and experience, present in all areas of life. A life-defining conflict arises when this fundamental bipolar tension turns into a conflictual polarization. An individuation-dependency conflict is present only if this constellation is of existential importance and formative for a person's life history: This conflict involves the activation of experiences that either seek or avoid closeness, not the shaping of relationships in terms of caregiving or avoiding caregiving. The theme of individuation-dependency deals with the question of being alone or the ability to be with others. In its pathological conflict version, it concerns the necessity of being alone or being with others as an existential requirement.

Task: Based on this context, classify the following interview excerpt regarding the theme of "autonomy-dependency" into one of the following categories: *"not present"*, *"not assessable"*, *"of little significance"*, *"significant"*, *"very significant"*. For tasks where the interviews were summarized prior to classification, the model was first instructed to generate a summary of the interview based on a provided example. This example was derived from the OPD Manual.

A.3 Classification Classes

- **Not assessable** – The category cannot be determined due to insufficient or ambiguous information. There may be a lack of relevant content, unclear statements, or methodological limitations preventing a reliable assessment.
- **Not present** – There is no indication of the characteristic or phenomenon being evaluated. The available information does not support its existence or relevance in the given context.
- **Of little significance** – The characteristic or phenomenon is present but plays only a minor role. It appears occasionally but does not have a substantial influence on behavior, emotions, or interactions.
- **Significant** – The characteristic or phenomenon is clearly identifiable and has a notable impact. It influences thoughts, emotions, or interactions and is relevant to the overall assessment.
- **Very significant** – The characteristic or phenomenon is a dominant feature. It strongly shapes experiences, interactions, or coping mechanisms and is central to the evaluation.

A.4 Extra Plots

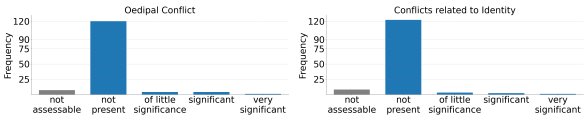

Figure 5: Class distributions for Oedipal Conflict and Conflicts related to Identity