# When Consistency Becomes Bias: Interviewer Effects in Semi-Structured Clinical Interviews

**Hasindri Watawana**[1,2], **Sergio Burdisso**[1], **Diego A. Moreno-Galván**[3]
**Fernando Sánchez-Vega**[3], **A. Pastor López-Monroy**[3], **Petr Motlicek**[1,4],
**Esaú Villatoro-Tello**[1]

[1] Idiap Research Institute, Switzerland
{hwatawana,sburdisso,pmotlicek,evillatoro}@idiap.ch

[2] EPFL, Switzerland

[3] Centro de Investigación en Matemáticas (CIMAT), Mexico
{diego.moreno,fernando.sanchez,pastor.lopez}@cimat.mx

[4] Brno University of Technology, Czech Republic

**Abstract**

Automatic depression detection from doctor–patient conversations has gained momentum thanks to the availability of public corpora and advances in language modeling. However, interpretability remains limited: strong performance is often reported without revealing what drives predictions. We analyze three datasets—ANDROIDS, DAIC-WOZ, and E-DAIC—and identify a systematic bias from interviewer prompts in semi-structured interviews. Models trained on interviewer turns exploit fixed prompts and positions to distinguish depressed from control subjects, often achieving high classification scores without using participant language. Restricting models to participant utterances distributes decision evidence more broadly and reflects genuine linguistic cues. While semi-structured protocols ensure consistency, including interviewer prompts inflates performance by leveraging script artifacts. Our results highlight a cross-dataset, architecture-agnostic bias and emphasize the need for analyses that localize decision evidence by time and speaker to ensure models learn from participants' language.

**Keywords:** Depression Corpora, Clinical Interviews, Graph Convolutional Network (GCN)

## 1. Introduction

Language has long been established as a powerful indicator of personality, socio-emotional state, and mental health (Pennebaker et al., 2003; Tackman et al., 2019). This insight has spurred a rich body of work at the intersection of AI, natural language processing, and clinical psychology, showing that structured interviews and written responses can reveal important aspects of cognitive and behavioral functioning, particularly for automatic depression detection (Malandrakis and Narayanan, 2015; Villatoro-Tello et al., 2021a,b).

While a growing body of recent work trains automatic depression detection models on both participant responses and interviewer prompts (Zhuang et al., 2024; Agarwal and Dias, 2024; Milintsevich et al., 2023; Shen et al., 2022), it remains unclear how much each contributes to model performance. This study asks: *to what extent can a model classify a participant as depressed using only the interviewer's questions?* Surprisingly, across three datasets (ANDROIDS, DAIC-WOZ, E-DAIC) and two model families, *interviewer-only* (*I*) models often match or outperform *participant-only* (*P*) models. This does not imply that interviewer questions are clinically more informative; instead, qualitative analyses show that *I*-models exploit systematic shortcuts tied to the interview script, focusing on recurring prompts and question positions—a phenomenon we term *prompt-induced bias*.

Our main contributions are threefold: (1) We quantify prompt-induced bias across multiple datasets, showing that *I*-models can outperform *P*-models. (2) We show that this effect is model-agnostic by reproducing it with both graph convolutional networks and transformers. (3) We provide qualitative analyses that reveal how *I*-models concentrate on narrow interview segments while *P*-models distribute evidence more broadly.

These findings highlight the importance of carefully handling interviewer prompts and verifying that models truly leverage participant language rather than spurious cues.

## 2. Datasets

Within clinical practice, the initial assessment of mental illness is a semi-structured interview in which clinicians pose a standardized yet open-ended sequence of questions to elicit symptoms, history, and functioning. This protocol balances consistency with flexibility, enabling reliable assessment while allowing patients to elaborate in their own words. The depression corpora we study are designed to simulate such standardized



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 2355–2361
11–16 May 2026. 

screening protocols for identifying people at risk: an interviewer (human or virtual) delivers a controlled set of prompts to ensure replicability and coverage of diagnostic cues, while participants respond freely, producing language that can be analyzed for clinical indicators.

### 2.1. DAIC-WOZ

The Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) dataset (Gratch et al., 2014) contains semi-structured clinical interviews in North American English conducted by "Ellie", an animated virtual interviewer controlled by a human in another room ("Wizard-of-Oz" setup where a user interacts with a mock interface controlled, to some degree, by a person). DAIC-WOZ is multimodal—audio, video, and manual transcripts—and includes PHQ-8 depression ratings for each participant (Kroenke et al., 2009). Ellie's design emphasizes replicability and consistency: she draws from a finite repertoire of 191 prompts spanning general questions (e.g., lifestyle and sleep), neutral backchannels, positive/negative empathic responses, surprise tokens, continuation prompts (e.g., "could you tell me more?"), and miscellaneous control items. This controlled prompting is intended to elicit behaviors associated with depression and related symptoms while reducing variability attributable to the interviewer. DAIC-WOZ comprises 189 subjects split into 107 train, 35 development, and 47 test interviews.

### 2.2. E-DAIC

E-DAIC (DeVault et al., 2014) is an extension of DAIC-WOZ that preserves the same interview format while scaling the collection. The key difference is that the virtual interviewer used in E-DAIC is fully automatic in contrast to the human-controlled Ellie in DAIC-WOZ. While DAIC-WOZ distributes complete two-speaker transcripts (interviewer and participant), E-DAIC provides only the participant side. Therefore, for our experiments, we prepare automatic transcripts for E-DAIC using WhisperX (Bain et al., 2023) pipeline. The process is explained in Section 3. We corrected several mislabeled subjects after verifying inconsistencies between PHQ scores and the provided binary labels, an issue noted previously by others (Ali et al., 2025). The official split sizes are 163 train, 56 development, and 56 test interviews.

### 2.3. ANDROIDS

ANDROIDS (Tao et al., 2023) is an Italian speech corpus for depression detection collected "in the wild" using laptop microphones. Each participant was recorded in two tasks: a Reading Task (RT)—everyone reads the same short, simple story to reduce literacy/education effects—and an Interview Task (IT) with spontaneous speech, where the interviewer is instructed to ask only minimal prompts, yielding semi-structured but low-intervention interactions. In this work we focus only on the data from the IT task. It comprises 116 native-Italian participants (64 depressed, 52 controls), with controls matched to the depressed cohort by gender, age, and education to minimize demographic confounds. Interviews are manually segmented into turns; diagnostic labels ("depressed" vs. "control") come from clinicians following DSM-5 criteria.[1] The release includes audio and acoustic features, but no ground-truth transcripts. Therefore, we generated WhisperX transcripts for ANDROIDS (see Section 3).

## 3. Methodology

**Data Preparation** Our study is based on the text modality. DAIC-WOZ provides complete transcripts including both interviewer and participant utterances. ANDROIDS doesn't provide ground-truth transcripts, and E-DAIC releases transcripts for the participant side only. For ANDROIDS and E-DAIC, we therefore built complete textual data as follows: using participant utterance timestamps provided in the metadata, we derived complementary (non-overlapping) interviewer timestamps, extracted interviewer and participant audio clips separately using the timestamp data, and generated automatic transcripts with the WhisperX ASR pipeline (Bain et al., 2023) (Whisper large-v3 (Radford et al., 2022) with a faster-whisper backend).[2] To have a fair comparison between the impact of using interviewer vs. participant utterances on automatic depression detection, we do not use the E-DAIC participant gold transcripts and instead rely on matched ASR transcripts for both speakers.

**Models** We evaluate two architectures to test if prompt effects are model-agnostic: (i) Longformer (Beltagy et al., 2020), a transformer with sparse attention suitable for long documents such as interviews, and (ii) GCN (Burdisso et al., 2023), a graph-based model with word and document nodes. Using these two types of models allows us to analyze results from both a contextualized, semantic transformer and a keyword-focused GCN.

- **Longformer:** A linear classification head is

[1] Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (Association, 2013).

[2] This pipeline yields a WER of 15.3% on the E-DAIC participant side, evaluated against the originally provided transcripts as ground truth.

| Model | Source | | ANDROIDS | | | DAIC-WOZ | | | E-DAIC† | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *P* | *I* | *Avg* | *D* | *C* | *Avg* | *D* | *C* | *Avg* | *D* | *C* |
| (Burdisso et al., 2023) | ✓ | | – | – | – | 0.84 | 0.80 | 0.89 | 0.80 | 0.67 | 0.94 |
| (Ilias and Askounis, 2024) | ✓ | ✓ | 0.93 | – | – | – | – | – | – | – | – |
| (Borraccino, 2025) | ✓ | ✓ | 0.92 | 0.93 | 0.9 | – | – | – | – | – | – |
| (Milintsevich et al., 2023) | ✓ | ✓ | – | – | – | 0.81 | – | – | – | – | – |
| *P*-Longformer | ✓ | | 0.79 | 0.82 | 0.76 | 0.71 | 0.61 | 0.81 | **0.67** | **0.56** | 0.79 |
| ***I***-Longformer | | ✓ | **0.98** | **0.98** | **0.98** | **0.73** | **0.64** | **0.83** | 0.65 | 0.50 | **0.80** |
| *P*-GCN | ✓ | | 0.93 | 0.95 | 0.92 | 0.85 | 0.81 | 0.88 | 0.70 | 0.54 | 0.86 |
| ***I***-GCN | | ✓ | **0.97** | **0.97** | **0.98** | **0.88** | **0.85** | **0.91** | **0.74** | **0.57** | **0.90** |

Table 1: Development-set $F_1$ scores on ANDROIDS, DAIC-WOZ, and E-DAIC. For each dataset, we report macro-average (*Avg*), Depressed (*D*), and Control (*C*) $F_1$. ANDROIDS scores are reported as 5-fold averages; DAIC-WOZ and E-DAIC use the official development split. Sources: participant (*P*) and interviewer (*I*) text. Upper block: text-only baselines from prior work. **Bold** = best in group; **underlined** = best overall text-only result. †: E-DAIC prior-work results use gold participant transcripts and are not directly comparable.

added on top of the pre-trained Longformer-BERT to classify using the *[CLS]* token; both the encoder and head are fine-tuned on the training split. Experiments are conducted with both longformer-mini-1024 and longformer-base-4096 but only the best-performing configuration is reported.

• **GCN:** We use the two-layer $\omega$-GCN of Burdisso et al. (2023), which represents each corpus as a graph with word and document (interview) nodes. Representations evolve through three stages: (i) an initial one-hot layer, (ii) a latent embedding after the first convolution, and (iii) a two-dimensional output after the second convolution, corresponding to depressed and control probabilities. Because words and documents share the same embedding space, the final layer provides class probabilities for both interviews and individual words, offering an interpretability handle that identifies which words—and interviewer prompts—serve as discriminative evidence (see Section 5).

**Quantifying Interviewer Bias** Our approach builds on Burdisso et al. (2024), who analyzed prompt-induced bias within the DAIC-WOZ corpus. We extend this line of work in three ways. First, we generalize the analysis across multiple clinical interview corpora—ANDROID and E-DAIC—to test whether the interviewer-related bias persists beyond a single dataset. Second, while Burdisso et al. (2024) focused on development splits, we also quantify the impact of interviewer bias on held-out test data, providing a clearer estimate of its effect on reported model performance. Third, we evaluate the bias using automatically generated transcriptions, as manual interviewer transcripts are unavailable for ANDROID and only partially available for E-DAIC.

This setup allows us to assess whether the bias remains detectable under more realistic, automatically transcribed conditions. For each architecture, we train and evaluate two model variants:

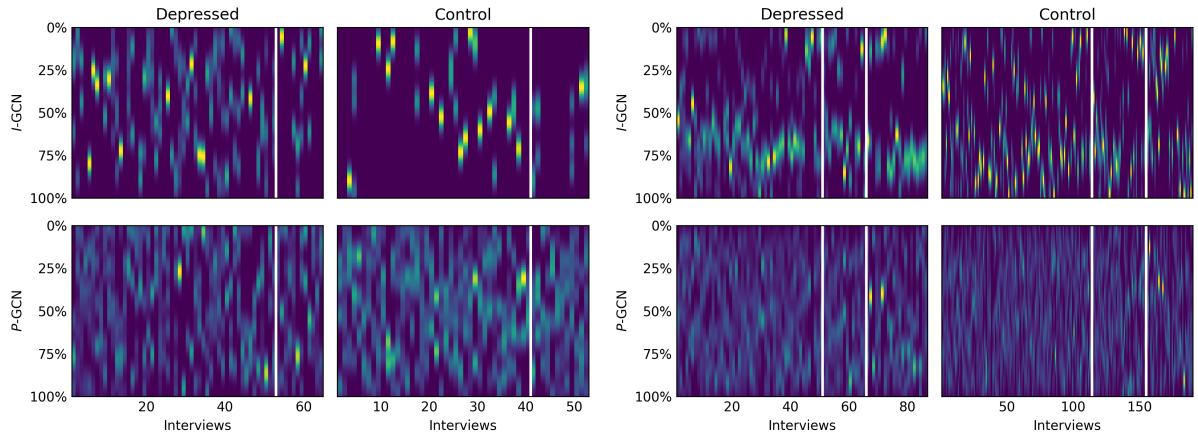
- *participant-only* (*P*): trained and evaluated using only the participant's responses
- *interviewer-only* (*I*): trained and evaluated using only the interviewer's prompts

These two model variants allow us to directly investigate our main research question by comparing performance when models have access only to participant responses (*P*) versus only to interviewer prompts (*I*). This comparison quantifies the extent to which interviewer questions carry implicit diagnostic information, revealing potential bias in automatic depression detection models.

## 4. Experiments and Results

Main results are reported in Table 1. For ANDROIDS five-fold averages are reported; for DAIC-WOZ and E-DAIC the official dev split is used. Across all three corpora we train two variants per architecture—participant-only (*P*-GCN and *P*-Longformer) and interviewer-only (*I*-GCN and *I*-Longformer). For each GCN model variant, we optimize and evaluate results using multiple setups (e.g. using all vocabulary, only top vocabulary etc), and record the best performance on dev split. Similarly for Longformer, best dev performance among longformer-mini-1024 and longformer-base-4096 is reported.

On DAIC-WOZ, the interviewer side is consistently stronger than the participant side for both model families (P-Longformer 0.71 → I-Longformer 0.73; P-GCN 0.85 → I-GCN 0.88).




(a) ANDROIDS (b) E-DAIC

Figure 1: Temporal heatmaps comparing keyword evidence learned by interviewer-only (*I*, top) vs. participant-only (*P*, bottom) models across interviews in the ANDROIDS and E-DAIC datasets. Each column represents one interview. The y-axis corresponds to the normalized interview timeline, where 0% marks the beginning of the interview and 100% marks its end. White vertical lines denote split boundaries (train/dev/test for E-DAIC; train/dev only for ANDROIDS). The ANDROIDS plot is shown for Fold 1.

| **Model** | **DAIC–WOZ** | | | **E–DAIC** | | |
|---|---|---|---|---|---|---|
| | *Avg* | *D* | *C* | *Avg* | *D* | *C* |
| *P*-Longformer | **0.68** | **0.54** | **0.82** | 0.40 | 0.17 | 0.62 |
| ***I***-Longformer | 0.53 | 0.27 | 0.78 | **0.56** | **0.44** | **0.68** |
| *P*-GCN | 0.59 | **0.56** | 0.63 | **0.54** | **0.47** | 0.62 |
| ***I***-GCN | **0.62** | 0.54 | **0.70** | **0.54** | 0.33 | **0.76** |

Table 2: Test-set $F_1$ scores (macro average *Avg*, *D*epressed, *C*ontrol) on DAIC–WOZ and E–DAIC, for participant-only (*P*) and interviewer-only (*I*) variants of Longformer and GCN. For each model, the test result corresponds to the same configuration that achieved the best dev performance in Table 1.

This reproduces prior findings reported in (Burdisso et al., 2024), that Ellie's prompts provide highly discriminative shortcuts, and shows the effect is not tied to a specific architecture.

On ANDROIDS, the bias is even more pronounced: I-Longformer achieves 0.98 macro-$F_1$ on dev gaining a 19% advantage over P-Longformer. I-GCN outperforms P-GCN by 4%. Despite being a different language and collection setting, interviewer prompts again dominate, reinforcing that the advantage arises from interview structure. This bias is evident on E-DAIC with the GCN models; with Longformer, the *P* and *I* variants perform comparably.

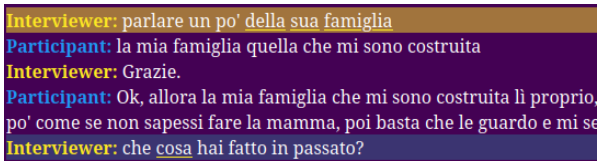
In Table 2 we also report results on the official test sets of DAIC-WOZ and E-DAIC to assess whether the interviewer bias carries over to unseen data. On E-DAIC, both GCN and Longformer models show that interviewer-only variants outperform or match participant-only variants, confirming that the bias persists beyond the dev split. On DAIC-WOZ, the effect is architecture-dependent: keyword-based GCNs still benefit from interviewer turns (I-GCN 0.62 vs P-GCN 0.59), while not in case of the context-based Longformer (P-Longformer 0.68 vs I-Longformer 0.53), suggesting that semantic modeling could partially mitigate prompt-driven shortcuts in some cases.

Overall, these results show that interviewer prompts can inflate performance in semi-structured interviews, independent of architecture. This cross-dataset contrast underscores our central point: gains attributed to "using interviewer questions as context" (Zhuang et al., 2024; Agarwal and Dias, 2024; Milintsevich et al., 2023; Shen et al., 2022) may simply reflect prompt-driven bias rather than improved modeling of the participant's language.

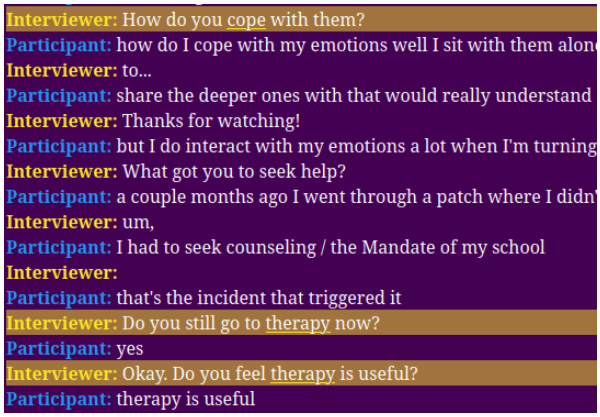
## 5. Analysis and Discussion

The heatmaps in Figure 1 visualizes where each model finds decision evidence (words used by the model to identify the depressed group; which we refer as "keywords") over time and by speaker. For every interview (x-axis), we aggregate the model's learned "keywords" along the normalized interview timeline (y-axis, 0–100%) and plot keyword density heatmaps separately for interviewer-only (*I*-GCN) and participant-only (*P*-GCN) models, and for Depressed vs. Control cohorts.

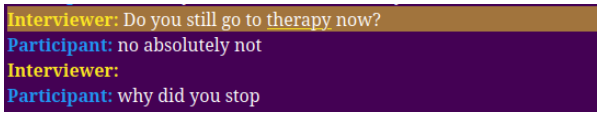
By analyzing the heatmaps across different datasets, we consistently found that patterns diverge sharply by speaker stream. *I*-GCN exhibits



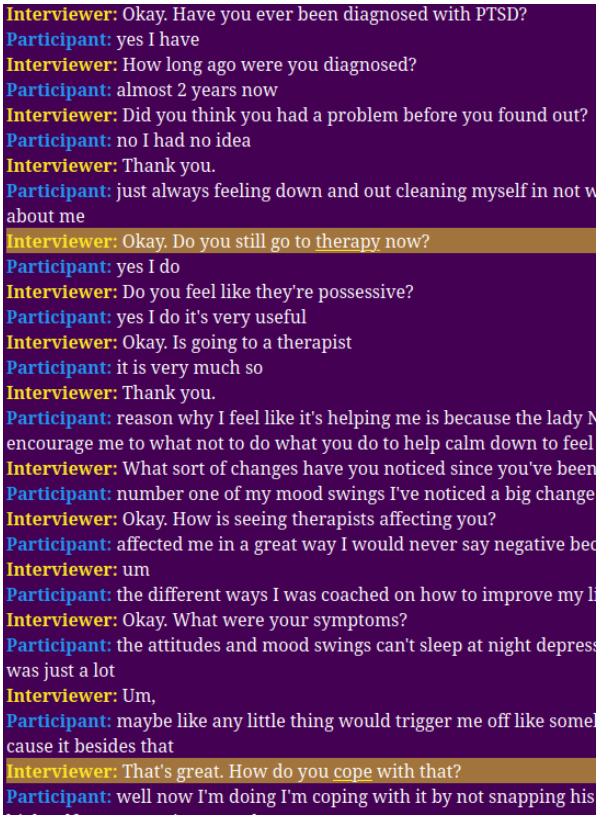

(a) Example interview from ANDROIDS. First utterance translates into 'talk about your family'




(b) Example of a depressed interview from E-DAIC




(c) Example of a control interview from E-DAIC




(d) Example of a depressed interview from E-DAIC

Figure 2: Color-coded interview excerpts in which prompts identified by the *I*-model as bias-carrying are highlighted. Underlined words denote the model's learned *keywords*, corresponding to the high-contrast narrow bands in Figure 1.

narrow, high-contrast bands, indicating that the model relies on specific interviewer turns to make its decision (concentrated keywords). In contrast, *P*-GCN shows low-contrast activity spread across most of the timeline, consistent with drawing evidence from many participant utterances rather than a few fixed positions. The qualitative content of those interviewer bands differs by dataset,

**ANDROIDS**: *I*-GCN repeatedly focus on prompts that probe family context, how the last week was spent, and work status (see Fig. 2a). These focused regions recur across interviews even though their absolute timing varies, suggesting the model has learned the type of prompt.

**E-DAIC and DAIC-WOZ**: *I*-GCN mainly concentrates on three prompts—"How do you cope with that?", "Do you still go to therapy?" followed by "Do you feel therapy is useful?"—while largely ignoring other clinically relevant prompts (e.g., PTSD screening, reasons for seeking help, symptoms, and effects of therapy). This again reflects selectivity for a small subset of interviewer turns. Depressed interviews in Fig. 2b and Fig. 2d clearly illustrate this shortcut behavior. Fig. 2c shows a control subject: after the highlighted bias prompt, the follow-up questioning pattern shifts, providing enough signal for the model to distinguish control from depressed interviews.

## 6. Conclusion

Our analysis shows that interviewer prompts enable models to distinguish between depressed and control participants across DAIC-WOZ, ANDROIDS, and E-DAIC, with Interviewer models focusing on specific, localized questions whereas Participant models distribute cues broadly across the conversation. These results highlight a key methodological issue: including interviewer turns introduces unintended biases, as models may exploit scripted prompts as shortcuts rather than learning genuine linguistic or behavioral markers of depression. Importantly, this bias is consistent across datasets and model architectures, underscoring its relevance as a general methodological concern rather than an artifact of any particular experimental setup. Future work should carefully account for the interviewer's influence when designing and evaluating conversational mental health assessment systems, for instance by isolating participant-only turns or developing bias-aware evaluation protocols.

## 7. Acknowledgments

This work was partially funded by the SNSF through the SPIRIT project **ORIENTER**: tOwards undeRstanding and modelIng the language of mENTal health disordERs (grant no. IZSTZ0_223488).

## 8. Ethics Statement

**Data privacy, consent, and dataset use.** Our study relies on publicly distributed clinical-interview corpora with documented consent and privacy safeguards. For DAIC–WOZ and E-DAIC, participants completed informed consent prior to interviews; consent materials included an option permitting data sharing for research. The released transcriptions underwent systematic de-identification (e.g., removal of names, specific dates, addresses) and identifying utterances are withheld; only appropriately anonymized audio/video features and transcripts are distributed under the institutional ethical guidelines, with broader raw data shared case-by-case. For ANDROIDS, data were collected in mental-health centers under institutional and national ethical regulations, with all participation voluntary and every participant signing an informed-consent letter; psychiatrists provided diagnostic labels (DSM-5 framework), and the release includes turn segmentation and person-independent protocols while protecting identity. In our work, we use text only (including ASR where ground truth is unavailable), do not attempt re-identification, and adhere to all dataset usage terms.

**Role of AI in Healthcare.** Our experiments are intended to underscore the value of interpretable, AI-assisted methods as decision *support*, not as replacements for clinicians. Diagnostic authority must remain with qualified professionals; delegating clinical decisions to an algorithm introduces unacceptable risk in high-stakes healthcare settings. By exposing shortcut learning and prompt-induced biases in interview data, our work contributes to the development of bias-aware models and evaluation practices for clinical interview analysis.

## 9. Limitations

Our study has several limitations.

**Transcript Quality and Ground Truth** ANDROIDS provides no manual transcripts and E-DAIC releases only participant transcripts; complete transcripts for both of these datasets were generated with ASR. Consequently, comparisons between *P* and *I* models can be affected by ASR errors and speaker segmentation, and any advantage/disadvantage may partly reflect transcription noise rather than underlying language alone. Ground-truth, two-speaker transcripts for E-DAIC and ANDROIDS would enable a more precise estimate of prompt-induced bias.

**Modality Restriction** Our analysis is text-only across ANDROIDS, DAIC-WOZ, and E-DAIC. While this isolates the linguistic contribution, it does not capture the acoustic and visual features, where interviewer structure and participant signals may manifest differently. As future work, we expect to extend the study to multimodal aspects, running the same *P* vs. *I* ablations to verify whether the prompt-induced bias persists, weakens, or strengthens when non-text modalities are included.

## 10. Bibliographical References

Navneet Agarwal and Gaël Dias. 2024. Analysing Relevance of Discourse Structure for Improved Mental Health Estimation. In *9th Workshop on Computational Linguistics and Clinical Psychology (CLPSYCH) associated to 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Saint Julian, Malta.

Abdelrahman A. Ali, Aya E. Fouda, Radwa J. Hanafy, and Mohammed E. Fouda. 2025. Leveraging audio and text modalities in mental health: A study of llms performance.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. American Psychiatric Publishing.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Annapia Borraccino. 2025. Modeling depressive patterns in italian discourse: Insights from natural language processing. Master's thesis, The Graduate Center, City University of New York (CUNY), New York, NY.

Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, Pastor López-Monroy, and Petr Motlicek. 2024.

Daic-woz: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. *arXiv preprint arXiv:2404.14463*.

Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, and Petr Motlicek. 2023. Node-weighted Graph Convolutional Network for Depression Detection in Transcribed Clinical Interviews. In *Proc. INTERSPEECH 2023*, pages 3617–3621.

Loukas Ilias and Dimitris Askounis. 2024. A Cross-Attention Layer coupled with Multimodal Fusion Methods for Recognizing Depression from Spontaneous Speech. In *Interspeech 2024*, pages 912–916.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Nikolaos Malandrakis and Shrikanth S Narayanan. 2015. Therapy language analysis using automatically generated psycholinguistic norms. In *Proc. Interspeech 2015*.

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

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE.

Allison M Tackman, David A Sbarra, Angela L Carey, M Brent Donnellan, Andrea B Horn, Nicholas S Holtzman, To'Meisha S Edwards, James W Pennebaker, and Matthias R Mehl. 2019. Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of personality and social psychology*, 116(5):817.

Esaú Villatoro-Tello, Gabriela Ramírez-de-la Rosa, Daniel Gática-Pérez, Mathew Magimai.-Doss, and Héctor Jiménez-Salazar. 2021a. Approximating the mental lexicon from clinical interviews as a support tool for depression detection. In *Proc. ICMI'21*, page 557–566.

Esaú Villatoro-Tello, S. Pavankumar Dubagunta, Gabriela Ramírez-de-la-Rosa Julian Fritsch, Petr Motlicek, and Mathew Magimai-Doss. 2021b. Late Fusion of the Available Lexicon and Raw Waveform-Based Acoustic Modeling for Depression and Dementia Recognition. In *Proc. Interspeech 2021*, pages 1927–1931.

Chen Zhuang, Deng Jiawen, Zhou Jinfeng, Wu Jincenzi, Qian Tieyun, and Minlie Huang. 2024. Depression detection in clinical interviews with LLM-empowered structural element graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

## 11. Language Resource References

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *Lrec*, volume 14, pages 3123–3128. Reykjavik.

Fuxiang Tao, Anna Esposito, and Alessandro Vinciarelli. 2023. The androids corpus: A new publicly available benchmark for speech based depression detection. In *Interspeech 2023*, pages 4149–4153.