Automated Detection and Classification of Delusion-related Content in Naturalistic Audio Diaries Using Multi-Agent Language Models

Feng Chen[1], Justin Tauscher[1], Changye Li[1],
Meliha Yetisgen[1], Alex Cohen[2], Adam Kuczynski[1],
Angelina Pei-Tzu Tsai[1], Benjamin Buck[3], Dror Ben-Zeev[1], Trevor Cohen[1]

[1]University of Washington, Seattle, WA
[2]Louisiana State University, Baton Rouge, LA
[3]University of North Carolina at Chapel Hill, NC
cohenta@uw.edu

Abstract

Speech monologues recorded in naturalistic settings provide opportunities to characterize mental illness phenomenology and detect symptom exacerbation. Large language models (LLMs) offer new possibilities for automating this process, as they require annotated data primarily for evaluation rather than training. In this paper, we present a novel automated, multi-agent LLM pipeline for the fine-grained, multi-label extraction of language suggestive of delusional beliefs, associated affective responses, and behavioral responses from transcripts of naturalistic audio diaries collected from people with moderate persecutory ideation. Evaluating an ensemble of three foundation models, we demonstrate that detailed diagnostic prompt instructions successfully reduce false positives for delusional theme classification, but also constrain the interpretation of affective or behavioral responses. Furthermore, comparing multi-agent adjudication frameworks shows that complex conversational debate between agents diminishes accuracy on clinically ambiguous text by inducing premature consensus. Instead, majority voting establishes robust performance (Micro F_1 of 0.872 and 0.779 for delusion detection and classification respectively). This work provides a validated and scalable pipeline for the automated detection and characterization of content suggesting delusional beliefs in naturalistic speech.

1 Introduction

Delusions are fixed, false beliefs that are held with strong conviction and are particularly resistant to counter-evidence (American Psychiatric Association, 2013). Although most commonly associated with psychotic disorders, attenuated forms of paranoid and persecutory thinking have also been reported in non-clinical populations, supporting the view that psychotic experiences occur along a continuum (Freeman et al., 2005; Heilskov et al., 2020; van Os et al., 2009). Delusions are multidimensional phenomena that vary in conviction and distress with content that often reflects recognizable thematic categories such as persecution, reference, grandiosity, and somatic concerns (Pappa et al., 2025). These experiences can be a major source of distress and functional impairment, strongly shaping how severely an individual's daily life is disrupted in the context of psychotic symptoms (Freeman, 2016; Buck et al., 2023; Kiran and Chaudhury, 2009).

Despite their clinical significance, systematically characterizing delusional content as it unfolds over time remains difficult. Current assessment approaches rely largely on structured clinician interviews conducted during intermittent clinical encounters, which may miss episodic fluctuations in delusional thinking and are vulnerable to recall bias and interpretive error (Haddock et al., 1999; Khan et al., 2013; Ben-Zeev et al., 2012). Capturing these fluctuations is critical for monitoring risk, guiding intervention, and understanding how delusional beliefs evolve over time (Bell et al., 2024).

Smartphone-based ecological momentary assessment (EMA) enables repeated symptom measurement in naturalistic settings (Myin-Germeys et al., 2018), and when paired with audio diary recordings, can capture rich first-person accounts of lived experience in real time (Ben-Zeev et al., 2020; Buck et al., 2023). When combined with automated speech recognition (Radford et al., 2022), these recordings can be transformed into large corpora of transcripts suitable for computational analysis. However, extracting clinically meaningful signals from large volumes of patient-generated language remains challenging, as manual review is time-consuming and difficult to scale (Tauscher et al., 2025; Wu et al., 2022).

Detecting presence of delusional content alone does not adequately characterize its clinical significance. Understanding its impact requires at-

Proceedings of the 11th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 12–31

tention to the thematic content of beliefs, the affective responses they evoke, and the behaviors they motivate. Delusional themes have been shown to relate to differences in etiology, distress, and need for treatment, while affective and behavioral responses to delusional beliefs are closely related to functional impairment and disability (Buck et al., 2023; Pappa et al., 2025). However, manually annotating these dimensions in large corpora of patient-generated language is time-consuming, costly, and difficult to scale, particularly given that inter-annotator agreement for psychiatric phenomena remains moderate even among trained raters (Tauscher et al., 2023; Di Forti et al., 2025). Large language models (LLMs) offer a potential solution. Their capacity for step-by-step chain-of-thought reasoning shows promise for mental health prediction tasks (Xu et al., 2024), and is well-suited to navigate the diagnostic logic and exclusion rules that fine-grained symptom characterization demands. However, individual models remain vulnerable to inconsistency and confabulation (Huang et al., 2023; Ji et al., 2023; Asgari et al., 2025). Multi-agent frameworks, in which multiple LLMs independently annotate and then adjudicate disagreements through voting, structured judging, or iterative debate (Du et al., 2023; Chan et al., 2023), offer a principled mechanism for resolving this ambiguity while retaining interpretable reasoning traces.

In this paper, we present an automated pipeline for multi-label annotation of delusional content in transcripts of audio diaries collected from individuals with delusions. We evaluate the approach based on data from the Study on Assessing Determinants and Antecedents of Persecutory Thoughts (ADAPT), which collected smartphone-based audio diaries from participants meeting the criteria for moderate persecutory ideation (Buck et al., 2023). We deploy three LLMs (GLM-4.6, GPT-OSS, and Qwen-3) and systematically compare three adjudication strategies: majority voting, direct expert adjudication, and multi-agent conversational debate.

Our contributions are as follows:

- A systematic ablation of prompt context across four progressively layered levels for multi-label annotation of delusional theme, affective response, and behavioral response from naturalistic first-person speech transcripts.
- A systematic comparison of three multi-agent adjudication strategies for resolving LLM disagreement on clinically nuanced annotation tasks.
- An empirical evaluation across four levels of prompt complexity showing that progressively layered clinical context reduces false positive annotations while preserving diagnostic sensitivity.

2 Background

2.1 NLP for Psychotic Symptoms and Naturalistic Speech

NLP has increasingly been used to study language patterns associated with psychotic-spectrum disorders (Deneault et al., 2024; Hua et al., 2025a; Corcoran et al., 2020). Some early work focused primarily on clinical documentation extracted from electronic health records (Cohen et al., 2008, 2005; Irving et al., 2021). While valuable, such records reflect the clinician's interpretation of a patient's experience rather than the patient's own language, potentially obscuring the linguistic expression of beliefs that computational models aim to identify. As a result, recent work has increasingly examined patient-generated language, including from smartphone-based audio diaries, which capture individuals' experiences in their own words as they occur in everyday settings (Ben-Zeev et al., 2020). Recent work has applied NLP to identify disorganized thinking and cognitive biases in audio diaries (Xu et al., 2021; Tauscher et al., 2025), but the use of NLP to characterize delusional content remains largely underexplored. Unlike formal thought disorder, which NLP models quantify via semantic speech disruptions (Elvevåg et al., 2007), delusions represent disruptions in normative patterns of belief: recognizing them involves interpreting the beliefs expressed in language rather than whether this language is contextually consistent. This makes the task of automatically identifying delusional content more dependent on contextual interpretation and background knowledge than many traditional NLP applications.

2.2 Large Language Models for Complex Clinical Annotation

Artificial intelligence is playing a growing role in mental health research and practice (Ben-Zeev, 2026). Supervised machine learning depends on massive, expertly annotated datasets, which are

prohibitively expensive to acquire for fine-grained psychiatric symptoms (Wu et al., 2022). For text data, LLMs have emerged as a powerful alternative, demonstrating robust zero-shot (without any labeled examples) and few-shot (with only a few) classification capabilities (Labrak et al., 2024; Sivarajkumar et al., 2024). By employing Chain-of-Thought (CoT) prompting (Wei et al., 2022), LLMs can be instructed to explicitly articulate their diagnostic process, actively weighing inclusion criteria against exclusion rules (e.g., distinguishing a somatic abnormality from a nihilistic conviction). While domain-specific fine-tuned models often remain competitive for mental health classification tasks (Xu et al., 2024), CoT-prompted LLMs offer the advantage when there is limited task-specific training data, making them particularly suited to domains where large annotated corpora are unavailable.

While LLMs have rapidly expanded the scope of mental health NLP, recent scoping reviews identify persistent methodological limitations, including reliance on simulated clinical vignettes, restriction to narrow diagnostic categories, and poor performance on tasks requiring interpretation of highly subjective mental states (Le Glaz et al., 2021; Hua et al., 2025c,b). With delusions in particular, the field currently lacks validated prompt engineering frameworks capable of navigating the cascading decision processes required for hierarchical multi-label extraction from unstructured narratives.

2.3 Resolving Subjective Ambiguity with Multi-Agent Frameworks

Despite the promise of CoT reasoning, relying on a single LLM for complex clinical annotation is inherently brittle. Single-model inferences are prone to well-documented failure modes, including LLM hallucinations (generating text that is unrelated to the content of a prompt) and sycophancy (Ji et al., 2023; Sumita et al., 2024). In psychiatric contexts, where a narrative may simultaneously exhibit features of multiple delusional themes (e.g., a government conspiracy [Persecutory] broadcast through a television [Reference]), individual models risk inconsistently collapsing toward a single dominant label.

To mitigate these concerns, the NLP community has adopted multi-agent frameworks. Ensemble methods such as majority voting effectively smooth out individual biases (Wang et al., 2023). More advanced architectures enable models to interact: *multi-agent debate* prompts LLMs to critique and iteratively revise each other's outputs to encourage divergent thinking (Du et al., 2023; Liang et al., 2024), while the *LLM-as-a-Judge* paradigm employs a large highly capable model to adjudicate conflicting outputs based on reasoning paths (Zheng et al., 2023).

Recent NLP work has begun to question the benefits of multi-agent debate: Choi et al. (2025) show across seven benchmarks that majority voting alone accounts for most reported gains, while Wynn et al. (2025) attribute debate degradation to model sycophancy, the tendency of models to defer to others rather than defending correct answers. However, these findings have been established primarily on objective benchmarks. Whether they extend to clinical ambiguity, where disagreement stems from nuanced psychiatric interpretation rather than verifiable factual error, remains underexplored. In this paper, we extend these emerging findings to the clinical domain by deploying and comparing multi-agent adjudication strategies on a multi-label annotation task involving naturalistic psychiatric narratives.

3 Methods

Figure 1 illustrates the overarching architecture of our end-to-end multi-agent clinical extraction pipeline, with four primary phases: naturalistic data collection, progressive prompt engineering, independent LLM inference, and multi-agent adjudication.

3.1 Dataset

We used transcripts from ADAPT (Buck et al., 2023), a smartphone-based study in which 231 individuals with at least moderate persecutory ideation from 43 US states completed 30 days of audio diary recordings capturing momentary experiences and beliefs in naturalistic settings. Recordings were transcribed to yield free-text narratives from this clinically heterogeneous sample.

For this study, we selected the first available diary transcript for each participant that was longer than 30 seconds. We retained only entries exceeding three sentences to ensure sufficient content for multi-label annotation, yielding a final corpus of **136 transcripts**. Of these, 14 were reserved as a development set for iterative prompt engineering and model configuration, while the remaining 122 constituted the held-out evaluation set for all final

Figure 1: The end-to-end multi-agent clinical delusion annotation pipeline.

performance assessments reported in this paper.

3.2 Annotation Guideline and Human Labeling

We developed a structured annotation guideline for labeling audio diary transcripts capturing three dimensions of delusion expression: (1) presence and thematic category of delusional belief, (2) affective responses expressed in relation to the belief, and (3) behavioral responses that reflect how individuals act upon these experiences. The schema was developed iteratively using meta-analytic syntheses of delusional themes cross-referenced with Diagnostic and Statistical Manual of Mental Disorders (DSM) descriptions of delusional disorders and SNOMED CT terminology (Pappa et al., 2025; American Psychiatric Association, 2013; SNOMED International, 2024). The affective categories were informed by Ekman's theory of basic emotions, and the behavioral response categories were informed by cognitive models of delusions that emphasize safety behaviors, avoidance, threat monitoring, and help-seeking (Tracy and Randles, 2011; Ekman, 1992; Freeman, 2016). A comprehensive listing of all target classes and their final clinical definitions is provided in Appendix Table A.3.

Annotation was conducted by two licensed clinical psychologists, with discrepancies resolved by an adjudicator with expertise in clinical annotation guideline development. Annotators completed structured training sessions using expert-labeled examples, followed by independent pilot coding of 10% of transcripts to assess guideline clarity and initial inter-rater consistency. Discrepancies identified during the pilot phase were reviewed with annotators and used to refine coding procedures before full annotation proceeded. The adjudicator reviewed all disagreements and assigned final labels based on guideline criteria, retaining multiple labels when clinically appropriate. Pre-adjudication agreement reached 63.3% for affective response and 61.8% delusion presence, with behavioral response at 60.0% as reported in Appendix Table A.1.

3.3 Prompt Design and Engineering

The foundational prompt (**Level 1**) includes only the instructions of the basic task, the required output template (see the Appendix Table A.4), and the names of the clinical categories. Subsequent levels progressively layered additional context from the guideline: **Level 2** adds formal clinical definitions for each target category; **Level 3** integrates exclusionary rules (e.g., "Do NOT use when...") and diagnostic heuristics (e.g., "Key test: Is there a threat?") to help models distinguish between highly similar categories like *Reference* and *Persecutory* delusions; and **Level 4** adds short, clinician-validated transcript excerpts mapped to specific labels to provide in-context learning for conversational register and tone. This incremental design allowed a systematic evaluation of how structured clinical guidance influences classification behavior. Appendix Table A.2 illustrates this progression using the *Avoidance/ Withdrawal* behavioral category as an example.

3.4 Independent LLM-Based Clinical Annotation

To automate the annotation pipeline and establish reliable baseline measurements, we deployed three state-of-the-art, open-weight Large Language Models (LLMs) with 4-bit quantization (Dettmers and Zettlemoyer, 2023): **GLM-4.6 (357B)** (Team et al., 2025), **GPT-OSS (120B)** (OpenAI, 2026), and **Qwen-3 (235B)** (Yang et al., 2025). All models

were securely hosted on private servers behind an institutional firewall. Each transcript was initially independently annotated by these three foundation models. During inference, each model was instructed to generate its internal "thinking process" before yielding the final structured JSON output, providing the explicit context required for downstream adjudication of the models' clinical reasoning.

All models are configured with greedy decoding, setting `temperature=0`, `top_k=1`, and `max_new_tokens=4096`. In rare instances where a model failed to generate a valid JSON object, the pipeline triggered a fallback execution with the limit extended to `8192` tokens. On the 14-transcript development set, this deterministic configuration produced *Delusion Presence* F_1 scores of 1.00 (GLM-4.6) and 0.857 (GPT-OSS), compared with means of 0.857 ($\pm$0.051) and 0.914 ($\pm$0.032) across five stochastic runs at temperature 0.7. We adopt the deterministic setting for full-corpus evaluation to ensure reproducibility.

3.5 Multi-Agent Adjudication Frameworks

We developed and systematically compared three distinct multi-agent adjudication strategies to resolve diagnostic ambiguity:

Tri-Model Majority Voting: We implemented a consensus approach where a final transcript label was determined by a simple majority vote across all three independent models. In rare instances of a complete three-way disagreement (where each model produced a unique categorical label), Qwen-3 was utilized as the definitive decision-maker due to its strong standalone alignment with manual annotations on the validation set.

Direct Expert Adjudication: To evaluate the efficacy of the LLM-as-a-Judge paradigm, Qwen-3 was strictly deployed as an expert judge to resolve discrepancies between the GLM-4.6 and GPT-OSS annotators. The judge model was provided with the original transcript, expert guidelines, and conflicting GLM and GPT results. This input included the internal thinking processes of both annotator models, allowing the judge to determine the most accurate classification based on logical adherence to the clinical criteria (detailed prompt is in the Appendix Table A.5).

Multi-Agent Conversational Adjudication: To test whether interactive debate encourages error correction on subjective text, GLM-4.6 and GPT-OSS were instructed to deliberate over their disagreements. Each agent defended its original annotation using clinical reasoning, with instructions to concede to the opposing model if its logic was more accurate, or to propose a combined multi-label solution. After a maximum of two conversational turns, Qwen-3 reviewed the entire deliberation history to issue a final, authoritative ruling (detailed prompt is in the Appendix Table A.6).

3.6 Multi-Label Agreement Evaluation

Clinical targets in this study, specifically *Delusion Themes*, *Affective Expression*, and *Behavioral Responses*, frequently involve multiple overlapping labels within a single transcript. Furthermore, the distribution of these thematic labels is heavily imbalanced. To evaluate system performance across these complex categorical spaces, we utilized multi-label metrics depending on the evaluation phase:

Inter-Model Reliability: We report inter-model reliability between different LLMs using both **micro-averaged** and **macro-averaged Cohen's Kappa (κ)**. Micro-averaged κ is computed on flattened binary indicator matrices across all label-instance pairs, while macro-averaged κ is the unweighted mean of per-class κ values, treating each category equally regardless of its frequency.

Performance Against Human Ground Truth: We utilized two complementary F_1 metrics to evaluate performance against manual annotations. **micro-averaged F_1** aggregates all label-level predictions across the corpus, providing a robust measure of the system's global classification accuracy. Conversely, **Example-based F_1** calculates performance independently for each transcript and averages the results, reflecting the practical reliability of the system on an individual diary-level basis.

Finally, we report *Delusion Presence*, a binary target derived from the *Delusion Type* annotations indicating whether any delusional theme was detected. This serves as a clinically relevant screening metric independent of the specific thematic categorization. Although our annotation guideline includes an affective intensity, we excluded this from primary performance analyses due to low inter-rater agreement (29%) among expert annotators on this dimension.

4 Results

4.1 Individual Model Performance against Human Ground Truth

As shown in Table 1, all models achieved reasonable performance at Level 1. Under the fully contextualized Level 4 prompt, micro-averaged F_1 for *Delusion Presence* was notably stable, with GLM-4.6 (0.857), GPT-OSS (0.873), and Qwen-3-235B (0.839) all consistently identifying clinically relevant content. This indicates that detecting the basic presence of content suggesting a delusion requires relatively little in-context instruction.

However, moving to the complex multi-label classification of *Delusion Type* demonstrated the utility of the structured prompt engineering framework (see Figure 2). When provided with only the simplest prompt (Level 1), models achieved disproportionately high micro-averaged recall scores (0.75–0.89) but significantly lower micro-averaged precision scores (0.52–0.60), indicating many false positive category assignments. This gap narrowed substantially as clinical rules and few-shot examples were progressively introduced. With Level 4, micro-averaged precision for *Delusion Type* increased to 0.67–0.75, raising micro-averaged F_1 to 0.74–0.78 across models (for a comprehensive metric breakdown, see Appendix Table A.10).

Interestingly, the *Affective* and *Behavioral* targets revealed that prompt complexity does not universally improve performance across all clinical targets. For *Behavioral Response*, introducing the Level 4 rules successfully improved micro-averaged F_1 across all models (rising from 0.46–0.51 in Level 1 to 0.58–0.62 in Level 4), pulling models closer to the experts' clinical interpretations. In contrast, the highly constrained Level 4 prompt actually degraded performance on the *Affective Response* category, with micro-averaged F_1 dropping from 0.77–0.81 (Level 1) to 0.69–0.77 (Level 4).

4.2 Robustness of Tri-Model Majority Voting

As illustrated in Figure 3 and detailed in Appendix Table A.11, ensembling the three models systematically smoothed out individual model variance and established a robust performance ceiling. Under the fully contextualized Level 4 prompt, the Majority Vote achieved a 0.779 micro-averaged F_1 on *Delusion Type*, slightly exceeding the best individual model (GLM-4.6 at 0.777), and demonstrated remarkable stability across all prompt complexities. Even under the simplest Level 1 prompt (Category Only), the ensemble maintained a 0.746 micro-averaged F_1 on *Delusion Type*, substantially outperforming standalone models attempting the same task without rules (e.g., GPT-OSS at 0.649). With *Affective Response*, the micro-averaged F_1 score with majority vote was 0.839, exceeding that of the best individual model (0.813 for Qwen-3).

Other targets also benefited from this stabilization. For *Behavioral Response*, the ensemble successfully leveraged the strict Level 4 rules to increase from a baseline 0.545 (Level 1) to a peak 0.640 micro-averaged F_1. For *Affective Response*, the ensemble plateaued around 0.800 under Level 4, further confirming that rigid clinical rules cannot fully resolve affective nuance, even across multiple agents.

4.3 The Limitations of Expert Adjudication and Debate

We isolated the performance of the Qwen-3-235B judge approaches using the fully contextualized Level 4 prompt. We stratified the 122 manual transcripts into two subsets: cases where the two primary annotators (GLM-4.6 and GPT-OSS) initially agreed, and cases where they disagreed. When GLM-4.6 and GPT-OSS independently agreed on the *Delusion Type* ($N = 91$), their micro-averaged F_1 against the human gold standard was exceptionally high (0.857), substantially outperforming Qwen's independent baseline (0.800) on the same subset. This confirms that when LLMs naturally converge, their consensus is highly trustworthy. However, the 31 transcripts where GLM and GPT diverged represent the most clinically ambiguous narratives in the dataset. On this challenging subset, individual micro-averaged F_1 dropped (GLM: 0.615, GPT: 0.603, shown in Table A.12).

Prompting Qwen-3 to act as an LLM adjudicator over the GLM and GPT reasoning traces did not yield meaningful improvements (see Appendix Table A.12). The *Direct Judge* achieved a 0.618 micro-averaged F_1 on *Delusion Type* for the disagreement subset, offering scant improvement over the individual baseline models. For the targets of *Affective Response* (0.590) and *Behavioral Response* (0.571), the Direct Judge actually *underperformed* relative to individual baseline models.

More strikingly, the *Multi-Agent Conversational Adjudication* framework actively diminished performance on this ambiguous subset. Instead of refining the diagnostic boundaries through debate,

Clinical Target	Prompt Level	Micro-averaged F_1					
		Individual Models			Ensemble	Adjudication†	
		GLM-4.6	GPT-OSS	Qwen-3	Majority Vote	Direct Judge	Debate
Delusion Presence	Level 1 (Category)	0.865	0.855	0.809	0.853	—	—
	Level 4 (Full Prompt)	0.857	**0.873**	0.839	0.872	0.866	0.835
Delusion Type	Level 1 (Category)	0.707	0.649	0.667	0.746	—	—
	Level 4 (Full Prompt)	0.777	0.762	0.736	**0.779**	0.775	0.765
Affective Response	Level 1 (Category)	0.797	0.774	0.813	**0.839**	—	—
	Level 4 (Full Prompt)	0.694	0.760	0.768	0.800	0.738	0.719
Behavioral Response	Level 1 (Category)	0.507	0.456	0.513	0.545	—	—
	Level 4 (Full Prompt)	0.615	0.582	0.582	**0.640**	0.584	0.634

Table 1: Comprehensive performance comparison (micro-averaged F_1) evaluated against the full test corpus ($N = 122$). The best overall performance for each target is highlighted in bold. †Adjudication scores reflect corpus-level performance: for transcripts where GLM-4.6 and GPT-OSS initially agreed ($N = 91$), their consensus is retained; the adjudicator is invoked only on disagreement cases ($N = 31$). See Table A.12 for stratified results.

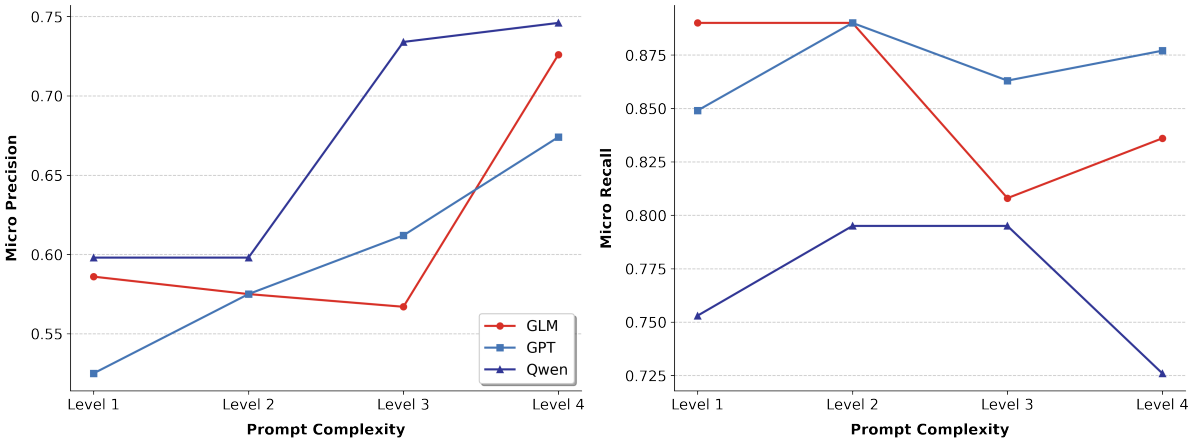

Figure 2: Micro-averaged precision (left) and recall (right) against human expert annotations for *Delusion Type* across the four prompt complexity levels.

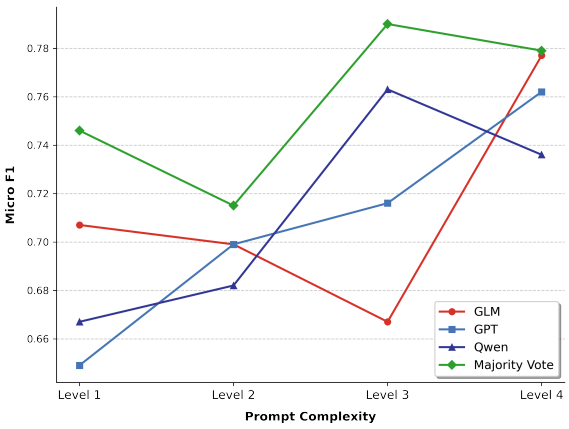

Figure 3: Micro-averaged F_1 for the Three Independent Models and Tri-Model Majority Vote against human expert annotations for *Delusion Type* across the four prompt complexity levels.

conversational adjudication reduced performance on *Delusion Type* down to a 0.545 micro-averaged F_1. This degradation was even more severe on the high-stakes binary screening task of *Delusion Presence*, dropping to 0.667.

A representative debate failure illustrates this pattern. In one transcript, the speaker states "I feel like people…Google me or research me online trying to find things to use against me," framed with uncertainty markers ("I feel like," "I guess") and self-awareness about the thought ("I have to take a while before I realize how irrational that they are"). The human annotators labeled this as *Persecutory*. GPT-OSS independently produced the correct *Persecutory* label, while GLM-4.6 rejected the presence of any delusion on the grounds that expressed doubt and insight are incompatible with fixed conviction. When the two models were placed in debate, GPT-OSS abandoned its correct

initial assessment within a single turn ("I agree with Annotator 1 – the passage does not contain a delusion"), restated GLM's form-versus-content reasoning, and explicitly conceded by Turn 2 ("I concede to Annotator 1's conclusion"). The judge confirmed the empty label, overriding GPT-OSS's correct initial annotation. This case illustrates how conversational adjudication can convert a correct prediction into an incorrect consensus: rather than defending an evidence-based position, the model with the right answer accommodated the more elaborately argued wrong answer.

4.4 Inter-Model Agreement and Prompt Complexity Analysis

To assess baseline consistency, we computed pairwise Cohen's κ across the three LLMs under the Level 4 prompt (see Appendix Table A.13). Models showed strong micro-averaged agreement on *Delusion Presence* (κ = 0.67–0.72) and *Delusion Type* (κ = 0.73–0.78). However, macro-averaged κ, which weights each class equally rather than allowing dominant labels to drive the score, was substantially lower for *Delusion Type* (κ = 0.43–0.52, averaged over 16 classes), reflecting weaker agreement on rare categories. This pattern was even more pronounced for response-related targets: *Affective Response* dropped from micro κ = 0.64–0.69 to macro κ = 0.30–0.41 (7 classes), and *Behavioral Response* from micro κ = 0.24–0.42 to macro κ = 0.09–0.20 (8 classes).

5 Discussion

To our knowledge, this study presents the first automated, multi-agent LLM pipeline for the fine-grained, multi-label annotation of delusions from naturalistic audio diaries. Without explicit constraints (Level 1), models generate many overlapping labels, capturing the correct category but maximizing recall at a severe expense to precision. This behavior aligns with the well-known tendency of LLMs to make broad and generalized semantic associations when faced with ambiguity (Ji et al., 2023; Huang et al., 2023). The introduction of strict exclusionary rules (Level 4) transformed this into a decision process with precise categorical boundaries. This finding is consistent with prior work on instruction-following, where including explicit negative examples and exclusionary guidance (e.g., "Things to Avoid") in task instructions has been shown to improve model generalization (Mishra et al., 2022; Wang et al., 2022). In our clinical context, negative constraints, rules specifying what *not* to label, were as important as category definitions for achieving precise classification.

While rigid definitions improved the extraction of delusional themes and behavioral responses, they actively impaired classification of affective responses. In naturalistic audio diaries, affect is rarely explicitly stated; it is implicit and highly contextual (Ritunnano et al., 2022). Strict exclusionary rules may pull models away from the intuitive consensus that human raters naturally share (Tauscher et al., 2023). This may reflect differences in the maturity and specificity of the annotation schema and prompt design across clinical targets rather than a fundamental limitation of prompt engineering. The affective response guidelines in this study were less extensively specified than the delusion-type taxonomy, potentially constraining the effectiveness of rigid rule-based prompting for these categories. Class imbalance likely compounds this effect: both affective and behavioral labels are dominated by a single category (*Fear-Anxiety* and *Avoidance/Withdrawal*, respectively; Appendix Tables A.8 and A.9), allowing simple prompts to achieve high scores through majority-class accuracy while strict rules introduce errors on less frequent categories.

Our findings on multi-agent adjudication further reveal that clinical ambiguity resists resolution through more elaborate reasoning and debate. On the 31 transcripts in which models disagreed, conversational debate degraded performance rather than resolving ambiguity, while majority voting remained the most robust strategy. We attribute the failure of multi-agent debate to a phenomenon where models prioritize agreement over accuracy, a behavior often referred to as model sycophancy (Sharma et al., 2023; Perez et al., 2023). In tasks with verifiable answers, debate can be effective, but in psychiatric evaluation there is seldom indisputable ground truth. When critiquing others' assessments of ambiguous audio diaries, the LLMs dispensed with their diagnostic prompts. Rather than defending their initial assessments, they were prone to excessive accommodation, frequently blending their initial assignments to reach a compromise.

Despite the inherent challenges of multi-label classification, our findings suggest that LLMs have strong potential for the rapid, automated assessment of diagnostic features of naturalistic narra-

tives. Across all evaluated models, prompt complexities, and adjudication frameworks, the baseline detection of *Delusion Presence* remained remarkably stable, consistently achieving a Micro F_1 exceeding 0.85. Notably, the ensemble's advantage was most pronounced on affective and behavioral targets: for *Affective Response*, majority voting at Level 1 achieved the highest score in our evaluation (0.839), substantially exceeding any individual model, suggesting that independent ensembling is particularly effective where individual models diverge. This indicates that our pipeline is a viable high-throughput tool for screening audio diaries, with the potential to flag clinically relevant content for expert review. Furthermore, by employing a simple tri-model majority vote alongside our Level 4 prompt, researchers can now automate the fine-grained extraction of specific delusional themes with a degree of reliability that rivals human inter-rater consensus (Di Forti et al., 2025). This approach addresses the traditional bottleneck of manual coding, presenting new avenues for longitudinal symptom tracking, and LLM-derived labels could further serve as training data for smaller, efficient classifiers deployable in resource-constrained settings.

6 Conclusion

In this work, we introduced a fully automated, multi-agent LLM pipeline for multi-label annotation of delusions, affect, and behavioral responses from unstructured audio diaries. Results from systematically comparing prompt engineering strategies indicate that while detailed clinical prompts successfully reduce confabulated false positive categories for descriptions of delusional content, they can constrain the interpretation of implicit affective or behavioral states. Furthermore, our evaluation of multi-agent adjudication frameworks shows that for clinically ambiguous text, complex interactive debate actively degrades model accuracy by inducing premature consensus. Instead, simply pooling the independent inferences of a heterogeneous LLM ensemble establishes robust performance. These findings demonstrate that LLM-based pipelines can reliably automate the characterization of delusional content in naturalistic speech, offering a scalable path toward continuous, longitudinal symptom monitoring that could support early detection of clinical deterioration and more timely intervention for individuals with serious mental illness.

Limitations

Several limitations of this study should inform future work. First, our evaluation was conducted on a relatively specialized corpus of audio diary transcripts collected from individuals experiencing moderate-to-severe persecutory ideation. Consequently, the dataset is heavily skewed toward *Persecutory* themes, and the micro-averaged F_1 scores reported here reflect agreement on frequent labels more than on rare categories. Future validation on broader psychiatric cohorts is necessary to confirm the generalizability of the pipeline. Second, while we evaluated three highly capable, distinct LLMs to ensure heterogeneity, the rapid evolution of model architectures creates the possibility that future foundation models may exhibit different baseline sycophancy behaviors. Third, the lower performance on affective and behavioral response classification likely reflects both the implicit and contextual nature of these targets and limitations in the current prompt design. Future work should explore alternative prompting strategies for these targets, such as dedicated annotation passes or less constrained prompt architectures that better accommodate the implicit nature of affective expression in naturalistic speech.

Ethical Considerations

Human subjects: The ADAPT study was approved by the University of Washington Institutional Review Board (IRB) under protocol STUDY00001321. All participants provided informed consent for audio recording and analysis (Buck et al., 2023). All transcripts were fully de-identified, and all LLMs were deployed locally to ensure no clinical narratives left the secure research environment.

Intended use and risks: This pipeline is a *research tool* for scalable annotation, not a diagnostic instrument. Automated classification of delusional content carries risks of misclassifications; any clinical application must include human expert review. Transcript examples used in the prompt templates are clinician-authored illustrations. The transcript excerpt presented in Section 4.3 is drawn from de-identified participant transcript.

Generalizability and reuse: Performance may vary across demographic or clinical populations beyond the ADAPT sample. Our prompt templates are intended for research purposes only and will be made available upon request.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) under award number UF1MH135901. The ADAPT dataset used in this study was collected with support from the National Institute of Mental Health (NIMH) under award number R01MH112641, and a NARSAD Young Investigator Grant from the Brain and Behavior Research Foundation (Grant #27917). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NIMH, or the Brain and Behavior Research Foundation.

References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition edition. American Psychiatric Association.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 8(1):274.

Imogen H. Bell, Emily Eisner, Stephanie Allan, Sharla Cartner, John Torous, Sandra Bucci, and Neil Thomas. 2024. Methodological Characteristics and Feasibility of Ecological Momentary Assessment Studies in Psychosis: a Systematic Review and Meta-Analysis. *Schizophrenia Bulletin*, 50(2):238–265.

Dror Ben-Zeev. 2026. To AI, or Not to AI: That Is Not the Question. *Psychiatric Services*, 77(2):94–95.

Dror Ben-Zeev, Benjamin Buck, Ayesha Chander, Rachel Brian, Weichen Wang, David Atkins, Carolyn J. Brenner, Trevor Cohen, Andrew Campbell, and Jeffrey Munson. 2020. Mobile RDoC: Using Smartphones to Understand the Relationship Between Auditory Verbal Hallucinations and Need for Care. *Schizophrenia Bulletin Open*, 1(1):sgaa060.

Dror Ben-Zeev, Gregory J. McHugo, Haiyi Xie, Katy Dobbins, and Michael A. Young. 2012. Comparing retrospective reports to real-time/real-place mobile assessments in individuals with schizophrenia and a nonclinical comparison group. *Schizophrenia Bulletin*, 38(3):396–404.

Benjamin Buck, Mary Wingerson, Justin S. Tauscher, Matthew Enkema, Weichen Wang, Andrew T. Campbell, and Dror Ben-Zeev. 2023. Using Smartphones to Identify Momentary Characteristics of Persecutory Ideation Associated With Functional Disability. *Schizophrenia Bulletin Open*, 4(1).

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. *arXiv preprint*. Version Number: 1.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. Debate or Vote: Which Yields Better Decisions in Multi-Agent Large Language Models? *arXiv preprint*. ArXiv:2508.17536 [cs].

Trevor Cohen, Brett Blatter, and Vimla Patel. 2005. Exploring dangerous neighborhoods: latent semantic analysis and computing beyond the bounds of the familiar. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2005:151–155.

Trevor Cohen, Brett Blatter, and Vimla Patel. 2008. Simulating expert clinical comprehension: adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative. *Journal of Biomedical Informatics*, 41(6):1070–1087.

Cheryl M. Corcoran, Vijay A. Mittal, Carrie E. Bearden, Raquel E. Gur, Kasia Hitczenko, Zarina Bilgrami, Aleksandar Savic, Guillermo A. Cecchi, and Phillip Wolff. 2020. Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*, 226:158–166.

Antoine Deneault, Alexandre Dumais, Marie Désilets, and Alexandre Hudon. 2024. Natural Language Processing and Schizophrenia: A Scoping Review of Uses and Challenges. *Journal of Personalized Medicine*, 14(7).

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pages 7750–7774. PMLR.

C. L. Di Forti, D. Liccione, and C. Scarpazza. 2025. Inter-rater reliability of psychiatric diagnosis: a systematic review and metanalysis. *European Psychiatry*, 68(Suppl 1):S191–S192.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint*. Version Number: 1.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Brita Elvevåg, Peter W. Foltz, Daniel R. Weinberger, and Terry E. Goldberg. 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93(1-3):304–316.

Daniel Freeman. 2016. Persecutory delusions: a cognitive perspective on understanding and treatment. *The Lancet Psychiatry*, 3(7):685–692.

Daniel Freeman, Philippa A. Garety, Paul E. Bebbington, Benjamin Smith, Rebecca Rollinson, David

Fowler, Elizabeth Kuipers, Katarzyna Ray, and Graham Dunn. 2005. Psychological investigation of the structure of paranoia in a non-clinical population. *The British Journal of Psychiatry: The Journal of Mental Science*, 186:427–435.

G. Haddock, J. McCARRON, N. Tarrier, and E. B. Faragher. 1999. Scales to measure dimensions of hallucinations and delusions: the psychotic symptom rating scales (PSYRATS). *Psychological Medicine*, 29(4):879–889.

Søren Esben Rytter Heilskov, Annick Urfer-Parnas, and Julie Nordgaard. 2020. Delusions in the general population: A systematic review with emphasis on methodology. *Schizophrenia Research*, 216:48–55.

Yining Hua, Suzanne V. Blackley, Ann K. Shinn, Joseph P. Skinner, Lauren V. Moran, and Li Zhou. 2025a. Identifying psychosis episodes in psychiatric admission notes via rule-based methods, machine learning, and pre-trained language models. *Translational Psychiatry*, 15(1):520.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, David A. Clifton, Andrew Beam, and John Torous. 2025b. Large Language Models in Mental Health Care: A Scoping Review. *Current Treatment Options in Psychiatry*, 12(1):27.

Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2025c. A scoping review of large language models for generative tasks in mental health care. *NPJ digital medicine*, 8(1):230.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv preprint*. Version Number: 2.

Jessica Irving, Rashmi Patel, Dominic Oliver, Craig Colling, Megan Pritchard, Matthew Broadbent, Helen Baldwin, Daniel Stahl, Robert Stewart, and Paolo Fusar-Poli. 2021. Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. *Schizophrenia Bulletin*, 47(2):405–414.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

Anzalee Khan, William Christian Yavorsky, Stacy Liechti, Guillermo DiClemente, Brian Rothman, Mark Opler, Ashleigh DeFries, and Sofija Jovic. 2013. Assessing the sources of unreliability (rater, subject, time-point) in a failed clinical trial using items of the Positive and Negative Syndrome Scale (PANSS). *Journal of Clinical Psychopharmacology*, 33(1):109–117.

Chandra Kiran and Suprakash Chaudhury. 2009. Understanding delusions. *Industrial Psychiatry Journal*, 18(1):3–18.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. 2024. A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2049–2066, Torino, Italia. ELRA and ICCL.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C. Ryan, Jonathan Marsh, Jordan DeVylder, Michel Walter, Sofian Berrouiguet, and Christophe Lemey. 2021. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research*, 23(5):e15708.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. *arXiv preprint*. ArXiv:2104.08773 [cs].

Inez Myin-Germeys, Zuzana Kasanova, Thomas Vaessen, Hugo Vachon, Olivia Kirtley, Wolfgang Viechtbauer, and Ulrich Reininghaus. 2018. Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry*, 17(2):123–132.

OpenAI. 2026. Introducing gpt-oss.

Elisavet Pappa, Fidelia Baah, Jessica Lynch, Lisha Shiel, Graham Blackman, Nichola Raihani, and Vaughan Bell. 2025. Delusional Themes are More Varied Than Previously Assumed: A Comprehensive Systematic Review and Meta-Analysis. *Schizophrenia Bulletin*, 51(3):637–645.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak

Supervision. *arXiv preprint*. ArXiv:2212.04356 [eess].

Rosa Ritunnano, Joshua Kleinman, Danniella Whyte Oshodi, Maria Michail, Barnaby Nelson, Clara S. Humpston, and Matthew R. Broome. 2022. Subjective experience and meaning of delusions in psychosis: a systematic review and qualitative evidence synthesis. *The Lancet. Psychiatry*, 9(6):458–476.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards Understanding Sycophancy in Language Models. *arXiv preprint*. Version Number: 4.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. *JMIR Medical Informatics*, 12:e55318.

SNOMED International. 2024. SNOMED CT.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. 2024. Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments. *arXiv preprint*. Version Number: 1.

Justin Tauscher, Xiruo Ding, Sarah Kopelovich, Arun Nagendra, Kevin Lybarger, Trevor Cohen, and Dror Ben-Zeev. 2025. Automated Flagging of Cognitive Biases in the Spoken Language of People with Hallucination Experiences. *Journal of Technology in Behavioral Science*.

Justin S. Tauscher, Kevin Lybarger, Xiruo Ding, Ayesha Chander, William J. Hudenko, Trevor Cohen, and Dror Ben-Zeev. 2023. Automated Detection of Cognitive Distortions in Text Exchanges Between Clinicians and People With Serious Mental Illness. *Psychiatric Services*, 74(4):407–410.

GLM-4.5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models.

Jessica L. Tracy and Daniel Randles. 2011. Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review*, 3(4):397–405.

J. van Os, R. J. Linscott, I. Myin-Germeys, P. Delespaul, and L. Krabbendam. 2009. A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness-persistence-impairment model of psychotic disorder. *Psychological Medicine*, 39(2):179–195.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint*. ArXiv:2203.11171 [cs].

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, and 16 others. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

Honghan Wu, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, Michael T. C. Poon, Natalie Fitzpatrick, Adam P. Levine, Luke T. Slater, Alex Handy, Andreas Karwath, Georgios V. Gkoutos, Claude Chelala, Anoop Dinesh Shah, Robert Stewart, Nigel Collier, Beatrice Alex, and 4 others. 2022. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *npj Digital Medicine*, 5(1):186.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. 2025. Talk Isn't Always Cheap: Understanding Failure Modes in Multi-Agent Debate. *arXiv preprint*. ArXiv:2509.05396 [cs].

Weizhe Xu, Jake Portanova, Ayesha Chander, Dror Ben-Zeev, and Trevor Cohen. 2021. The centroid cannot hold: comparing sequential and global estimates of coherence as indicators of formal thought disorder. In *AMIA Annual Symposium Proceedings*, volume 2020, page 1315.

XUHAI Xu, BINGSHENG YAO, YUANZHE DONG, SAADIA GABRIEL, HONG YU, JAMES HENDLER, MARZYEH GHASSEMI, ANIND K. DEY, and DAKUO WANG. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(1):31.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs.CL].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 46595–46623, Red Hook, NY, USA. Curran Associates Inc.

Appendix

Clinical Target	**N**	**Agree**	**Disagree**
Delusion Presence[a]	136	61.8%	38.2%
Delusion Type[a]	136	58.1%	41.9%
Delusion Type[b]	60	91.7%	8.3%
Affective Response[b]	60	63.3%	36.7%
Affective Intensity[c]	31	29.0%	71.0%
Behavioral Response[b]	60	60.0%	40.0%

Table A.1: Pre-adjudication agreement between two licensed clinical psychologist annotators. [a]Calculated across all transcripts. [b]Calculated on transcripts where both annotators agreed on delusion presence. [c]Calculated on transcripts where both annotators agreed on the presence of an affective response.

Complexity Level	Prompt Text Provided to LLM (Example: Avoidance/Withdrawal)
Level 1: Base	1. Avoidance/Withdrawal:
Level 2: + Definition	1. Avoidance/Withdrawal: Behaviors aimed at escaping, avoiding, or disengaging from situations, people, or cues linked to distress or delusional beliefs. The individual reduces contact rather than taking action to increase safety.
Level 3: + Rules	1. Avoidance/Withdrawal: Behaviors aimed at escaping, avoiding, or disengaging from situations, people, or cues linked to distress or delusional beliefs. The individual reduces contact rather than taking action to increase safety. Key test: Is the person pulling away, avoiding, or removing themselves from something?
Level 4: + Examples (Final Prompt)	1. Avoidance/Withdrawal: Behaviors aimed at escaping, avoiding, or disengaging from situations, people, or cues linked to distress or delusional beliefs. The individual reduces contact rather than taking action to increase safety. Examples: i. "I don't leave my house anymore because I'm scared someone will follow me." ii. "I avoid drinking tap water because I think it's poisoned." iii. "I stopped watching TV because I feel like they're sending me messages." Key test: Is the person pulling away, avoiding, or removing themselves from something?

Table A.2: Incremental prompt construction methodology, demonstrating how clinical context was progressively layered using the Avoidance/Withdrawal category.

Class	Definition
Clinical Target: Delusion Type	
Reference	Belief that neutral events, objects, or people contain special messages meant specifically for the individual.
Grandiosity	Belief that one possesses exceptional abilities, fame, power, identity, or destiny.
Religious	Beliefs involving divine authority, supernatural missions, spiritual transformation, or demonic forces.
Mind Reading	Belief that the individual can read other people's thoughts or mentally access their internal experiences.
Persecutory	Belief that others intend harm, surveillance, sabotage, or conspiracy against the individual.
Guilt/Sin	Belief that one has committed catastrophic wrongdoing or deserves extreme punishment despite no evidence.
Control	Belief that one's actions, emotions, or impulses are being controlled by an external force.
Somatic	Delusions about the body being diseased, damaged, infested, altered, or malfunctioning. Includes hypochondriacal delusions when medically plausible illnesses are believed with delusional certainty.
Nihilistic	Belief that oneself, parts of the body, or the world no longer exist, are dead, or have been destroyed. Includes a conviction that a major catastrophe will occur.
Erotomanic	Belief that another person, often someone of higher status, is in love with the individual.
Sexual	Delusional beliefs involving sexual content, intrusion, coercion, exposure, or sexual manipulation.
Jealous	Belief, without evidence, that a partner is being unfaithful.
Thought Withdrawal	Belief that an outside force is removing one's thoughts.
Thought Insertion	Belief that thoughts are placed into one's mind by an outside entity.
Thought Broadcasting	Belief that one's thoughts are being transmitted to others.
Unspecified	When the dominant delusional belief cannot be clearly determined or is not described in the specific types.
Clinical Target: Affective Response	
Fear-Anxiety	Indicates expressed fear, worry, tension, or apprehension about the experience.
Sadness-Despair	Indicates expressed sadness, hopelessness, emotional exhaustion, or despair.
Anger-Frustration	Indicates expressed irritation, anger, resentment, or frustration.
Euphoria-Excitement	Indicates expressed elation, pleasure, excitement, or energized positive affect (can include grandiose themes).
Hope-Optimism	Indicates expressed hopefulness, positive expectation, or belief in improvement or recovery.
Satisfaction-Contentment	Indicates expressed calm, peace, fulfillment, or a stable sense of well-being.
Neutral-None	Indicates no clearly expressed emotion, or a flat, matter-of-fact tone without affective coloring.
Clinical Target: Behavioral Response	
Avoidance/Withdrawal	Behaviors aimed at escaping, avoiding, or disengaging from situations, people, or cues linked to distress or delusional beliefs. The individual reduces contact rather than taking action to increase safety.
Safety-Seeking/Protective Behaviors	Behaviors intended to prevent harm or increase perceived safety, while still engaging with the environment. These include active protective steps, precautionary rituals, or hypervigilant routines.
Confrontation/Resistance	Behaviors involving direct action toward the perceived threat or source of delusional content such as challenging, disrupting, or fighting back.
Help-Seeking	Behaviors involving reaching out to others including friends, family, clinicians to obtain support, validation, or practical assistance.
Self-Soothing/Regulation	Behaviors aimed at calming oneself, reducing distress, or managing emotions internally. These are coping behaviors rather than actions directed toward external threats.
Engagement/Acceptance	Behaviors showing constructive engagement with reality, acceptance of experiences, or efforts to refocus on meaningful activity. Often reflects cognitive reframing or functional coping.
Risky or Harmful Behaviors	Behaviors that pose risk of harm to the individual or others, including self-harm, aggression, reckless acts, or substance misuse.
Neutral/None	Indicates that no specific behavioral response is described, or the individual does not mention any actions related to their experiences.

Table A.3: Comprehensive list of classes and their exact clinical definitions for Delusion Type, Affective Response, and Behavioral Response, as provided to the annotators and language models in the final guidelines.

Base Task Instructions and Output Formatting

ANNOTATION GUIDELINES FOR CLINICAL TARGETS

Audio diary transcripts will be annotated for various clinical targets, including hallucination content, delusions, substance use, and other relevant factors.

Each diary will be labeled as containing one or more clinical targets, and the relevant spans within the diary entry will be identified and annotated. These spans will represent triggers and entities associated with each clinical target, providing a detailed understanding of the individual's experiences and symptoms.

The following sections describe how each clinical target will be annotated. For each annotated clinical target, some entities are required and must always be present, while others may not always be present.

Note: A transcript may contain multiple delusion themes. These can be:

• The same span with multiple themes (e.g., a statement that is both Persecutory and Reference)

• Different spans with different themes

List each `delusion_span` and `delusion_type` pair separately in order.

Follow the following response template for your final answer.

Response Template and In-Context Formatting Examples

example input

I know they are monitoring my email, and I feel afraid all the time.

response template

`delusion_span`: "I know they are monitoring my email"
`delusion_type`: Persecutory
`affective_span`: "I feel afraid all the time"
`affective_category`: Fear-Anxiety
`affective_intensity`: Moderate
`behavioral_span`: null
`behavioral_category`: null

example input (multiple delusion themes)

God told me I am chosen to save humanity, and the government is trying to stop me.

response template

`delusion_span`: "God told me I am chosen to save humanity"
`delusion_type`: Religious
`delusion_span`: "God told me I am chosen to save humanity"
`delusion_type`: Grandiose
`delusion_span`: "the government is trying to stop me"
`delusion_type`: Persecutory
`affective_span`: null
`affective_category`: null
`affective_intensity`: null
`behavioral_span`: null
`behavioral_category`: null

example input (no delusion detected)

Today was a good day. I went to work and had lunch with a friend.

response template

`delusion_span`: null
`delusion_type`: null
`affective_span`: null
`affective_category`: null
`affective_intensity`: null
`behavioral_span`: null
`behavioral_category`: null

Table A.4: The high-level prompt structure (Base Task Instructions) provided to the language models. This foundational section establishes the extraction task, provides rules for handling multiple overlapping labels, and enforces a strict structured output format using explicit in-context examples.

Framework 1: Direct Adjudication Prompt (Example: Delusion Type)

You are an expert clinical judge evaluating delusion type annotations. Two models have annotated the same text and disagree on the delusion type.

Here are the original annotation guidelines that were used:
`{delusion_guidelines}`

— Original Text —
`{text}`

— MODEL A ANALYSIS —
Thinking Process: `{glm_thinking}`
Detailed Annotation: `{glm_response}`
Classification: `{glm_delusion}`

— MODEL B ANALYSIS —
Thinking Process: `{gptoss_thinking}`
Detailed Annotation: `{gptoss_response}`
Classification: `{gptoss_delusion}`

Based on the clinical guidelines above and considering both models' reasoning and annotation spans, determine the correct classification. You may:
- Choose Model A if their annotation is more accurate
- Choose Model B if their annotation is more accurate
- Choose Combined if both have valid points and the final answer should merge or differ from both

Provide your judgment in this exact format:
WINNER: (Model A | Model B | Combined)
REASONING: (Brief explanation)
CORRECT_TYPE: (The correct delusion type or comma-separated list if multiple)

Table A.5: System prompt used for the Direct Adjudication framework. In this single-pass approach, the judge model is provided with the independent reasoning traces of both original annotators and asked to select or synthesize a final clinical label.

Framework 2: Conversational Debate (Role Instructions & Discussion Prompt)

Annotator 1 Role: You are Annotator 1. Review the disagreement with Annotator 2 and either: Defend your original annotation with clinical reasoning; Concede to Annotator 2's annotation if you find their reasoning more accurate; Propose a combined solution if both annotations have merit. Provide brief clinical reasoning focused on the annotation guidelines.

Annotator 2 Role: You are Annotator 2. Review Annotator 1's response and either: Defend your original annotation with clinical reasoning... (same as Annotator 1).

Discussion Prompt:
You are a professional clinical annotator participating in an annotation discussion to resolve disagreements on delusion type classification.
`{delusion_guidelines}`

Original text: `{text}`

Annotation disagreement on: Delusion Type
Annotator 1's original annotation and reasoning:
- Delusion span: `{glm_delusion_span}`
- Delusion type: `{glm_delusion_type}`
- Original thinking: `{glm_thinking}`
- Original response: `{glm_response}`

Annotator 2's original annotation and reasoning:
- Delusion span: `{gpt_delusion_span}`
- Delusion type: `{gpt_delusion_type}`
- Original thinking: `{gpt_thinking}`
- Original response: `{gpt_response}`

`{conversation_history}`
`{current_role_instruction}`

Framework 2: Conversational Debate (Final Adjudicating Judge Prompt)

You are an expert clinical judge resolving an annotation disagreement on delusion type classification.
`{delusion_guidelines}`

Original text: `{text}`

(... Original Annotator 1 and Annotator 2 states injected here ...)

Discussion between annotators:
`{conversation_history}`

Based on the clinical guidelines, determine which annotator's ORIGINAL delusion type value is correct:
- If Annotator 1's original value (`{glm_delusion_type}`) is correct, choose Annotator 1
- If Annotator 2's original value (`{gpt_delusion_type}`) is correct, choose Annotator 2
- If neither is fully correct and a different/combined value is needed, choose Combined

Respond in this format:
Winner: (Annotator 1 or Annotator 2 or Combined)
Final delusion_type: (the correct value - must match winner's original value if Annotator 1 or 2, or new value if Combined)
Reasoning: (brief clinical justification)

Table A.6: System prompts used for the Conversational Debate framework. This pipeline utilizes an iterative, multi-turn discussion between the original annotator agents, guided by role instructions, before the entire debate history is passed to a final judge model for adjudication.

Delusion Type	GPT-OSS	GLM-4.6	Qwen-3	Manual
Persecutory	77	66	57	58
None	43	52	61	59
Reference	7	4	4	4
Control	4	3	2	2
Unspecified	1	3	1	3
Nihilistic	2	2	1	1
Thought Broadcasting	1	1	1	3
Religious	1	2	2	0
Somatic	1	2	1	1
Mind Reading	0	1	0	1
Grandiosity	0	0	2	0
Thought Insertion	1	0	0	0

Table A.7: Distribution of specific *Delusion Type* labels in the test set extracted by the three primary models under the fully contextualized Level 4 prompt setting, compared against manual expert annotations.

Affective Response	GPT-OSS	GLM-4.6	Qwen-3	Manual
Fear-Anxiety	73	61	71	51
Sadness-Despair	25	26	25	11
Anger-Frustration	14	13	9	8
Satisfaction-Contentment	1	1	1	0
Hope-Optimism	0	1	0	0

Table A.8: Distribution of specific *Affective Response* labels in the test set extracted by the three primary models under the fully contextualized Level 4 prompt setting, compared against manual expert annotations.

Behavioral Response	GPT-OSS	GLM-4.6	Qwen-3	Manual
Avoidance/Withdrawal	36	32	31	24
Help-Seeking	18	13	20	1
Safety-Seeking/Protective	9	8	11	4
Self-Soothing/Regulation	6	3	6	1
Confrontation/Resistance	3	8	4	1
Engagement/Acceptance	2	8	2	1
Risky/Harmful	5	3	1	1

Table A.9: Distribution of specific *Behavioral Response* labels in the test set extracted by the three primary models under the fully contextualized Level 4 prompt setting, compared against manual expert annotations.

Model	Prompt Level	Micro-averaged Precision	Micro-averaged Recall	Micro-averaged F1	Example F1
GLM-4.6	Level 1 (Category)	0.586	**0.890**	0.707	0.773
	Level 2 (Definitions)	0.575	**0.890**	0.699	0.743
	Level 3 (Rules)	0.567	0.808	0.667	0.711
	Level 4 (Full Prompt)	**0.726**	0.836	**0.777**	**0.811**
GPT-OSS	Level 1 (Category)	0.525	0.849	0.649	0.737
	Level 2 (Definitions)	0.575	**0.890**	0.699	0.775
	Level 3 (Rules)	0.612	0.863	0.716	0.785
	Level 4 (Full Prompt)	**0.674**	0.877	**0.762**	**0.798**
Qwen-3-235B	Level 1 (Category)	0.598	0.753	0.667	0.722
	Level 2 (Definitions)	0.598	**0.795**	0.682	0.771
	Level 3 (Rules)	0.734	**0.795**	**0.763**	**0.811**
	Level 4 (Full Prompt)	**0.746**	0.726	0.736	0.793

Table A.10: Comprehensive performance breakdown for *Delusion Type* extraction against human expert annotations.

Clinical Target	Prompt Level	Micro-averaged F1			
		GLM-4.6	GPT-OSS	Qwen-3-235B	Majority Vote
Delusion Presence	Level 1 (Category)	0.865	0.855	0.809	0.853
	Level 2 (Definitions)	0.836	0.871	0.837	0.868
	Level 3 (Rules)	0.822	0.886	0.868	0.878
	Level 4 (Full Prompt)	0.857	0.873	0.839	0.872
Delusion Type	Level 1 (Category)	0.707	0.649	0.667	0.746
	Level 2 (Definitions)	0.699	0.699	0.682	0.715
	Level 3 (Rules)	0.667	0.716	0.763	0.790
	Level 4 (Full Prompt)	0.777	0.762	0.736	0.779
Affective Response	Level 1 (Category)	0.797	0.774	0.813	0.839
	Level 2 (Definitions)	0.767	0.803	0.732	0.780
	Level 3 (Rules)	0.730	0.803	0.750	0.797
	Level 4 (Full Prompt)	0.694	0.760	0.768	0.800
Behavioral Response	Level 1 (Category)	0.507	0.456	0.513	0.545
	Level 2 (Definitions)	0.519	0.545	0.500	0.564
	Level 3 (Rules)	0.458	0.528	0.500	0.482
	Level 4 (Full Prompt)	0.615	0.582	0.582	0.640

Table A.11: Comprehensive performance comparison (Micro-averaged F1) across varying prompt complexities for individual LLMs and the Tri-Model Majority Vote.

	Agreement Subset ($N = 91$)		Disagreement Subset ($N = 31$)				
Clinical Target	**GLM/GPT Consensus**	**Qwen-3**	**GLM-4.6**	**GPT-OSS**	**Qwen-3**	**Direct Judge**	**Debate**
Delusion Presence	0.894	0.847	0.769	0.833	0.821	0.800	0.667
Delusion Type	0.857	0.800	0.615	0.603	0.612	0.618	0.545
Affective Response	0.850	0.800	0.409	0.612	0.711	0.590	0.500
Behavioral Response	0.667	0.375	0.606	0.567	0.635	0.571	0.629

Table A.12: Micro-averaged F1 performance stratified by initial annotator agreement. The Agreement Subset ($N = 91$) compares the natural consensus of GLM-4.6 and GPT-OSS against Qwen-3 as an independent baseline. All evaluations used the Level 4 prompt.

Clinical Target	GLM vs GPT		GPT vs Qwen		GLM vs Qwen	
	Micro	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**
Delusion Presence	0.709	0.709	0.672	0.672	0.721	0.721
Delusion Type	0.777	0.515	0.755	0.481	0.731	0.433
Affective Response	0.646	0.295	0.694	0.412	0.648	0.401
Behavioral Response	0.239	0.092	0.424	0.182	0.300	0.196

Table A.13: Pairwise agreement (Cohen's κ) across the three primary LLMs using the fully contextualized Level 4 prompt. Micro-averaged κ is computed on flattened binary indicator matrices; macro-averaged κ is the unweighted mean of per-class κ values. For *Delusion Presence* (binary), the two are equivalent.

	GLM-4.6 vs GPT-OSS		GPT-OSS vs Qwen-3		GLM-4.6 vs Qwen-3	
Clinical Target	**Agree**	**Disagree**	**Agree**	**Disagree**	**Agree**	**Disagree**
Delusion Presence	86.1%	13.9%	83.6%	16.4%	86.1%	13.9%
Delusion Type	75.4%	24.6%	73.8%	26.2%	76.2%	23.8%
Affective Response	65.6%	34.4%	66.4%	33.6%	65.6%	34.4%
Behavioral Response	47.5%	52.5%	58.2%	41.8%	53.3%	46.7%

Table A.14: Baseline pairwise agreement rates across the three primary LLMs using the fully contextualized Level 4 prompt. Agreement represents exact set matches; disagreement includes both partial overlap and zero intersection.