

Synthetic Depression Screening Dialogue Generator

A Multi-Agent Framework for Diverse, Realistic Patient-Doctor Conversations

Table of Contents

1. [Abstract](#)
2. [Introduction](#)
3. [Background and Related Work](#)
4. [Research Gaps and Design Philosophy](#)
5. [System Architecture](#)
6. [Patient Profile Generation](#)
7. [Doctor Configuration](#)
8. [Conversation Orchestration](#)
9. [Example Dialogues](#)
10. [Output and Data Structure](#)
11. [Current Limitations](#)
12. [Future Work](#)
13. [Applications](#)
14. [Usage Guide & Extending the System](#)
15. [References](#)
16. [Appendices](#)

1. Abstract

This project presents a **multi-agent framework** for generating synthetic doctor-patient conversations for depression screening. The system addresses a key limitation of existing approaches: producing *diverse, varied conversations* while maintaining *realism and clinical accuracy*.

Existing approaches struggle to achieve both goals simultaneously. Generated conversations often lack stylistic diversity despite different patient configurations, and assigned character personas tend to degrade toward generic patterns as dialogues progress.

Our framework addresses these challenges through **four core contributions**:

1. Separation of Behavioral Reasoning and Dialogue Generation

A dual-manager architecture separates the decision of *how* a character should behave (disclosure level, emotional state, response style) from the generation of natural speech. Manager agents reason about the full personality profile each turn, while conversational agents receive explicit behavioral constraints.

Key benefit: This separation prevents style drift, allows focused reasoning at each layer, and provides unambiguous guidance to dialogue generation.

2. Situationally-Responsive Patient Simulation

Patients respond to conversational context rather than following fixed scripts. Each turn, the system evaluates the doctor's approach, the patient's personality, and the conversation state to determine appropriate disclosure, emotional expression, and response calibration.

Key benefit: Disclosure evolves with trust-building, and agent emotional states shift in response to doctor behavior.

3. Compositional Diversity Through Structured Psychological Dimensions

The system defines multiple profile dimensions (Big Five templates, voice style axes, episode severity, life context) that combine multiplicatively. Weighted probability sampling across dimensions biases generation toward psychologically coherent combinations while preserving variation.

Key benefit: Produces thousands of unique but plausible individuals from small sets of building blocks.

4. Emergent Interactive Dynamics

Because both doctor and patient agents are responsive systems, interaction dynamics emerge without explicit scripting. Warm doctors gradually unlock guarded patients; dismissive doctors elicit increased guardedness; brisk doctors and verbose patients produce natural tension.

Key benefit: Conversation trajectories depend on both parties' behavior, creating realistic clinical dynamics.

Output: The system generates complete DSM-5 depression screening conversations with ground truth symptom labels, suitable for training machine learning models in mental health applications.

2. Introduction

2.1 The Problem

Machine learning models for mental health applications require training data. Real patient-doctor conversations present several acquisition challenges:

Challenge	Description
Access barriers	Privacy regulations, consent requirements, and clinical gatekeeping limit data availability
Cost	Professional transcription, annotation, and clinical review are resource-intensive
Limited diversity	Data from individual clinics reflects narrow patient populations

Synthetic data offers a potential solution if it meets two criteria:

- Sufficient **realism** for downstream utility
- Sufficient **diversity** to prevent model overfitting

2.2 Limitations of Existing Approaches

Existing LLM-based patient simulators face tradeoffs:

- **PatientSim** — Achieves clinical realism but struggles with nuanced language proficiency and communication styles
- **PATIENT-Ψ** — Provides rich cognitive modeling but includes no demographic variation
- **D4 and MDD-5k** — Criticized for limited stylistic variation across patient profiles
- **Demographic prompting** — Produces inconsistent results and risks generating stereotyped outputs

2.3 Our Approach

This work pursues diversity through **psychological dimensions** rather than demographic categories.

Research indicates that LLMs perform poorly when prompted to simulate demographic groups. We instead focus on dimensions that LLMs can reliably reproduce:

- Personality traits (Big Five model)
- Communication style (verbosity, directness, trust level)
- Emotional state and disclosure behavior
- Life context and stressor themes

These dimensions are *behaviorally grounded* and do not carry the cultural assumptions embedded in demographic categories.

2.4 Document Overview

This report provides a complete overview of the system:

1. Theoretical grounding in the literature
2. Architectural decisions and their justifications
3. Technical implementation details
4. Example outputs demonstrating diversity
5. Limitations and directions for future work

3. Background and Related Work

3.1 LLM-Based Patient Simulators

Several research groups have developed patient simulation systems for clinical training and data generation.

PatientSim targets Emergency Department settings, building patient profiles from MIMIC-IV clinical data across four persona axes:

- Personality/communication style
- Language proficiency
- Medical history recall
- Cognitive confusion

Evaluation: Clinician evaluations rated interactions as realistic, though language proficiency proved difficult to simulate accurately.

PATIENT-Ψ focuses on CBT training, developing cognitive models based on CBT concepts:

- Core beliefs
- Automatic thoughts
- Coping strategies

The system includes six conversational styles (plain, upset, verbose, reserved, tangent, pleasing). Expert evaluations rated cognitive accuracy highly. The system includes no demographic modeling.

D4 and MDD-5k are depression diagnosis datasets created through human annotation and LLM expansion. Reviewers noted limited stylistic variation, with conversations appearing similar despite different patient profiles.

PSYCHE targets psychiatric assessment with behavioral descriptors covering appearance, tone, verbal productivity, mood/affect, and insight. The system relies on generic "act consistent" instructions without systematic mechanisms to ensure trait fidelity.

3.2 Communication Style Control

Research on controlling LLM communication styles identifies consistent limitations.

Cohn et al. (2024) tested Satir communication styles (Appeaser, Blamer, Rationalizer, Distractor, Leveler) and found that LLMs approximate target styles in early turns but *drift toward neutral assistant patterns* as conversations continue. Style fidelity degrades over multi-turn dialogues.

Truong et al. (2025) evaluated persona-augmented generation using PersonHub personas—resume-style descriptions including occupation, interests, hobbies, attitudes, and demographics. They found that more challenging persona configurations consistently reduced model performance across tasks.

This quality-diversity tradeoff suggests that simply adding persona complexity does not improve output quality.

3.3 Sociodemographic Prompting

Attempts to add sociodemographic diversity through prompting have produced mixed results.

Beck et al. (2023) found that demographic cues (e.g., "Answer as a 25-year-old woman") change model behavior, but effects are inconsistent across models, tasks, and prompt variations. In some cases, prompting exacerbates stereotypes rather than representing groups authentically.

Sun et al. (2025) evaluated sociodemographic prompting for subjective judgments and found that zero-shot predictions align more closely with White annotators than Asian or Black annotators. Demographic prompts do not reliably improve alignment and sometimes worsen it.

Dialect generation research (Finch et al., 2025; Hofmann et al., 2024) demonstrates that LLMs produce exaggerated or stereotyped language when prompted for dialects. Users from dialect groups often prefer standard language in professional contexts.

3.4 Summary of Literature Findings

Finding	Source
Style drift over multi-turn dialogues	Cohn et al., 2024
Demographic prompting unreliable	Sun et al., 2025
Dialect simulation exaggerated	Finch et al., 2025
Quality-diversity tradeoff	Truong et al., 2025
Single-agent persona maintenance difficult	PatientSim evaluation

4. Research Gaps and Design Philosophy

4.1 Identified Gaps

The literature review identifies several gaps that existing systems do not address:

1. **No systematic integration** of communication style, cognitive state, and clinical content within a unified framework
 2. **No mechanism** for maintaining persona consistency across extended multi-turn dialogues
 3. **Demographic diversity approaches** produce stereotypes rather than authentic variation
 4. **Single-agent architectures** lack the interactive dynamics present in real clinical conversations
-

4.2 Design Decisions

Included Dimensions

Dimension	Rationale
Big Five personality templates	Grounded in validated psychological research; LLMs reliably manifest personality differences
Voice style dimensions	Controls <i>how</i> patients speak, independent of <i>what</i> they say
Age-appropriate contexts	Different life stages involve different stressors
Disclosure gradients	Trust develops over conversation; adds realistic dynamics
Multiple doctor personas	Doctor behavior affects patient responses

Excluded Dimensions

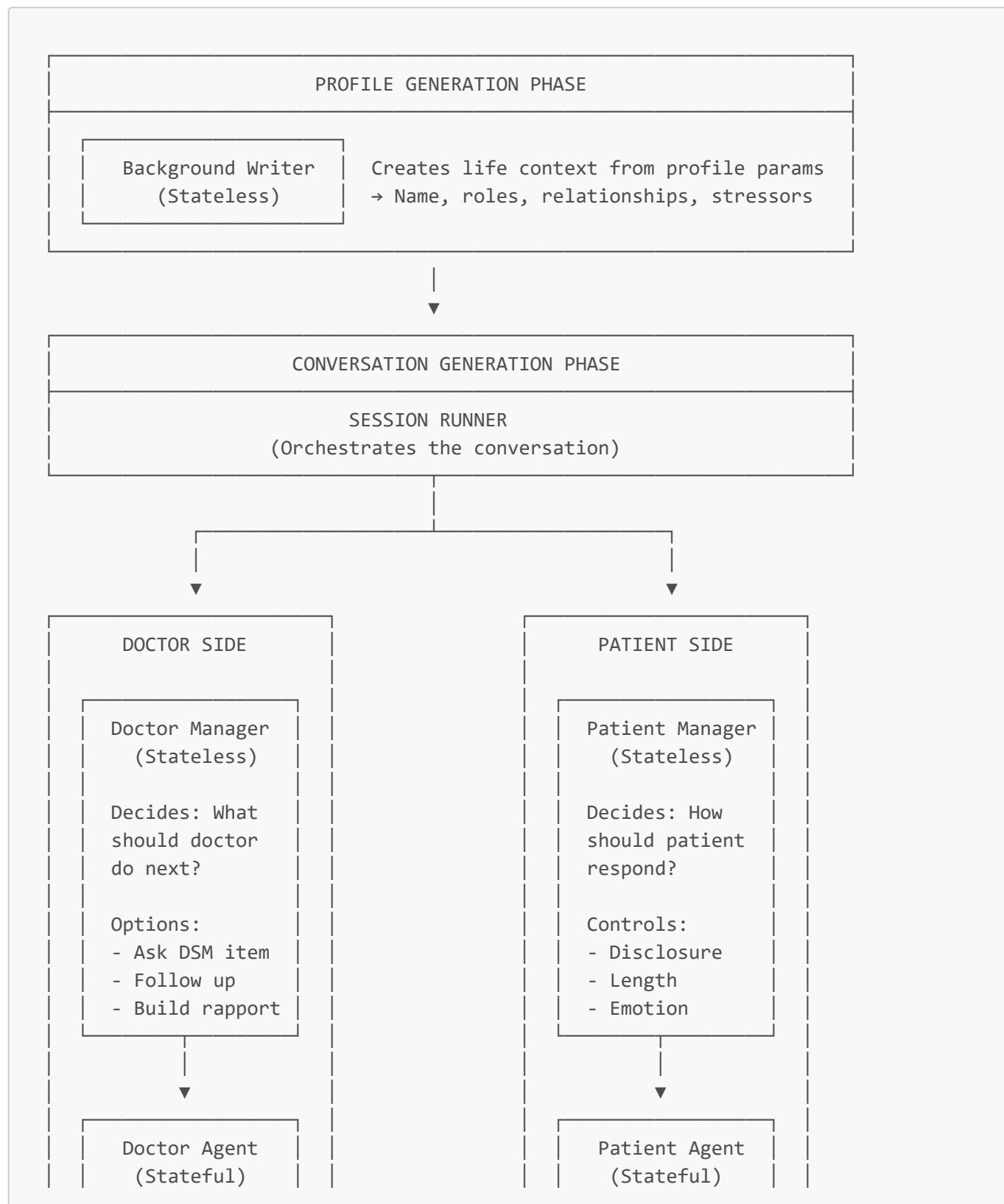
Dimension	Why Excluded
Race/ethnicity encoding	LLMs produce stereotyped outputs (Hofmann et al., 2024)
Dialect simulation	Unreliable and often exaggerated (Finch et al., 2025)
Explicit demographic stereotypes	"Act like a working-class person" produces caricatures
Socioeconomic status as behavior modifier	Introduces unpredictable biases (Sun et al., 2025)

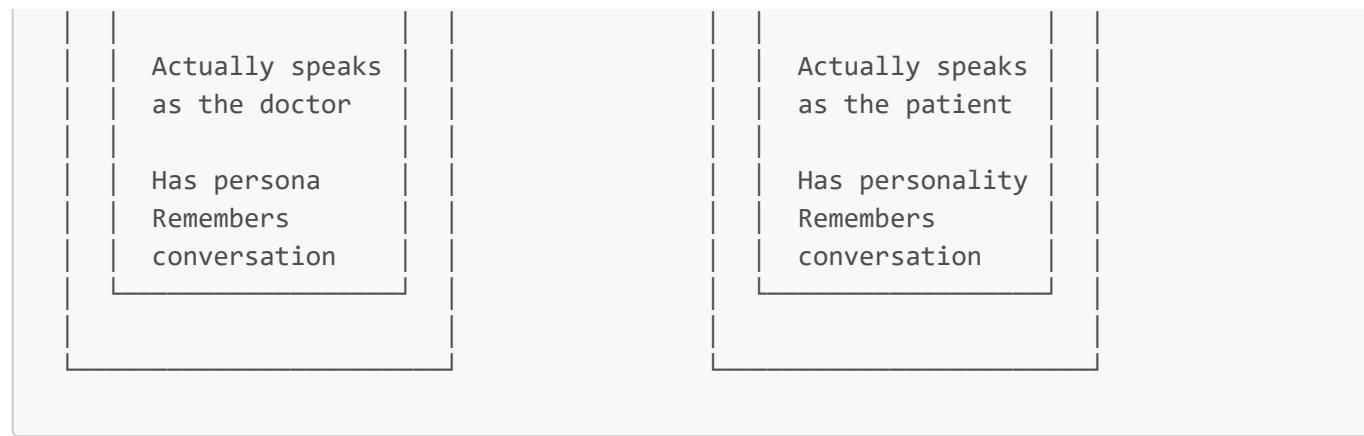
5. System Architecture

5.1 Overview

The system uses **five LLM agents** across two phases:

1. **Profile Generation Phase** — Background Writer creates patient life contexts
2. **Conversation Generation Phase** — Four agents (two manager-agent pairs) conduct screening dialogue





5.2 Manager-Agent Separation

This architecture addresses the **style drift problem** identified in the literature (Cohn et al., 2024). When a single agent handles both behavioral strategy and dialogue generation, it tends to drift toward generic patterns over extended conversations. Separating these functions prevents this degradation.

Manager Function

The manager determines *how* the character should behave given the current context:

- Evaluates disclosure level, emotional state, and response length
- Outputs structured guidance (JSON)
- Does *not* generate dialogue text

Agent Function

The agent receives explicit behavioral constraints from the manager:

- Generates natural dialogue within those constraints
- Focuses on naturalistic language production (word choice, phrasing, speech patterns)
- Does *not* make strategic decisions about disclosure or emotional presentation

Benefits of Separation

Benefit	Description
Focused reasoning	Each component handles a narrower task
Drift prevention	Manager re-evaluates full personality profile each turn
Explicit guidance	Manager provides parameters like <code>disclosure_stage: MINIMIZE</code>
Stylistic focus	Agent dedicates processing to speech patterns and vocabulary

After 20 turns, the manager still reasons from the complete profile rather than accumulated context—preventing the style degradation seen in single-agent approaches.

5.3 Complete Turn Flow

Each conversation turn proceeds through **four steps**:

SINGLE TURN FLOW

STEP 1: DOCTOR MANAGER DECIDES

Input:

- Full conversation history
- Remaining DSM symptoms to cover
- Doctor persona and microstyle
- Patient profile summary
- Turn budget status

Output (JSON):

- next_action: "DSM" | "FOLLOW_UP" | "RAPPORT"
- reason: why this action
- doctor_instruction: guidance for doctor agent
- dsm_symptom_key: which symptom (if DSM)



STEP 2: DOCTOR AGENT SPEAKS

Input:

- Directive tag with manager instruction
Example: <NEXT_QUESTION>
Ask about sleep patterns naturally...
</NEXT_QUESTION>
- Last patient message

Output:

- Natural dialogue as the doctor character



STEP 3: PATIENT MANAGER DECIDES

Input:

- Doctor's message and action type
- Patient personality profile
- Depression symptom ground truth
- Current disclosure state
- Full conversation history

Output (JSON):

- directness: "LOW" | "MED" | "HIGH"
- disclosure_stage: "MINIMIZE" | "PARTIAL" | "OPEN"
- target_length: "SHORT" | "MEDIUM" | "LONG"

- emotional_state: "neutral" | "tearful" | "frustrated" | etc.
- tone_tags: ["guarded", "self-critical", ...]
- key_points_to_reveal: what to mention
- key_points_to_avoid: what to hide
- patient_instruction: guidance for patient agent



STEP 4: PATIENT AGENT SPEAKS

- Input:**
- Manager guidance block
 - Doctor's message

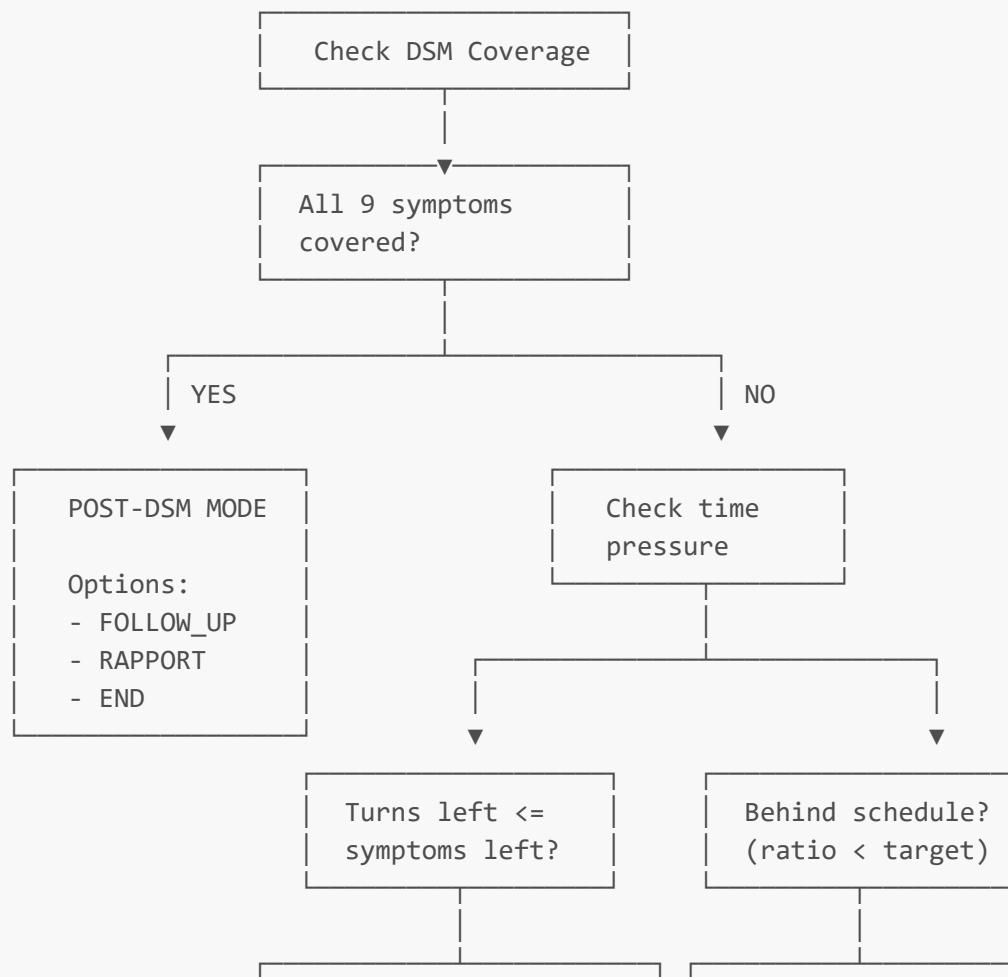
- Output:**
- Natural dialogue as the patient character

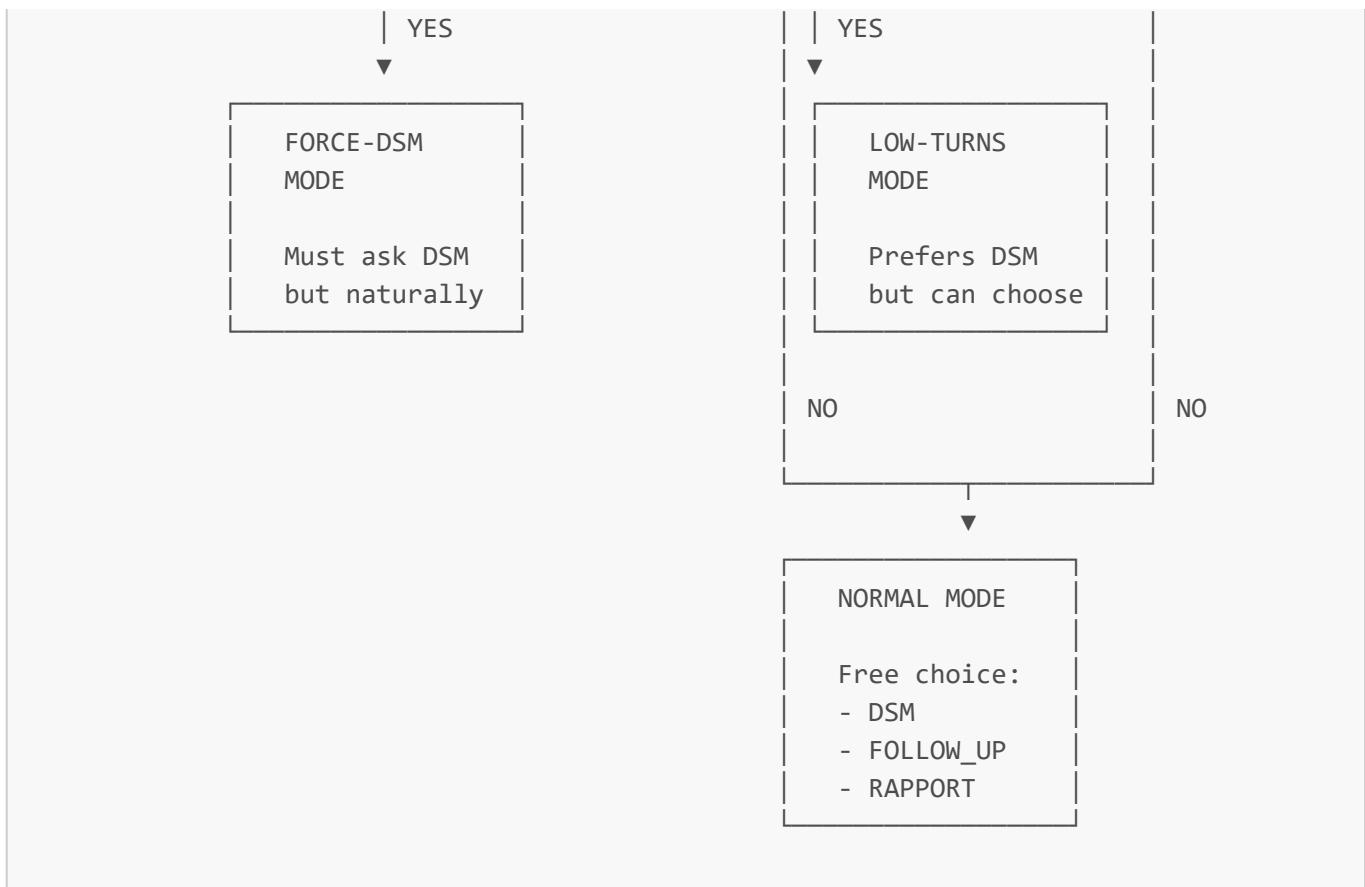


[Next turn begins]

5.4 Manager Mode Selection

The Doctor Manager operates in different modes depending on conversation progress:





6. Patient Profile Generation

Synthetic patients are constructed from **multiple independent dimensions** that combine to create unique individuals.

6.1 Big Five Depression Templates

The system uses **8 templates** based on how different Big Five personality extremes commonly manifest depression:

Template	Core Presentation	Emphasized Symptoms
NEUROTICISM_HIGH	Intense emotions, rumination, self-blame	Depressed mood, Guilt, Fatigue, Concentration
EXTRAVERSION_LOW	Withdrawal, flatness, low energy	Anhedonia, Psychomotor slowing, Fatigue
CONSCIENTIOUSNESS_HIGH	Perfectionism, controlled affect	Weight/appetite, Sleep, Fatigue, Guilt
CONSCIENTIOUSNESS_LOW	Apathy, disorganization	Psychomotor, Fatigue, Concentration
AGREEABLENESS_HIGH	Guilt about others, over-concern	Depressed mood, Sleep, Fatigue, Guilt
AGREEABLENESS_LOW	Irritability, external blame	Depressed mood, Psychomotor agitation, Sleep
OPENNESS_HIGH	Existential sadness, philosophical	Depressed mood, Anhedonia, Suicidal thoughts

Template	Core Presentation	Emphasized Symptoms
OPENNESS_LOW	Limited insight, somatic focus	Weight/appetite, Sleep, Fatigue

Each template includes **trait modifiers** that add texture: "worry-prone," "perfectionistic," "socially withdrawn," "self-blaming," etc. Zero to two modifiers are sampled per patient.

6.2 Voice Style System

Voice style controls *how* the patient communicates, independent of their personality content:

Dimension	Options	What It Controls
Verbosity	terse, moderate, detailed	Response length and elaboration
Expressiveness	flat, balanced, intense	Emotional display in language
Trust	guarded, neutral, open	Willingness to share with doctor
Intellect	low, moderate, high functioning	Vocabulary complexity, abstract thinking
Humor	none, occasional, frequent	Use of humor (often as coping mechanism)

These dimensions are sampled independently, creating **81+ unique voice combinations** ($3 \times 3 \times 3 \times 3 = 81$, plus humor variations).

6.3 Symptom Distribution Logic

Symptom assignment follows a **five-step process**:

Step 1: Template Selection

The Big Five template determines 3-4 "emphasized" symptoms that are most relevant to that personality type.

Step 2: Episode Density Selection

Density controls overall symptom severity:

Density	Description
ULTRA_LOW	Subclinical (0-2 symptoms total)
LOW	Mild presentation (mostly NONE/RARE)
MED	Moderate (mix of RARE/SOME/OFTEN)
HIGH	Severe (more SOME/OFTEN)

Step 3: Emphasized Symptom Assignment

For each emphasized symptom:

- Assign an intensity level (LOW/MED/HIGH)
- Sample frequency based on intensity:

Intensity	Frequency Distribution
LOW	50% RARE, 40% SOME, 10% OFTEN
MED	50% SOME, 40% OFTEN, 10% RARE
HIGH	20% SOME, 80% OFTEN

Step 4: Non-Emphasized Symptom Assignment

For remaining symptoms, use density-based weights:

Density	Frequency Distribution
LOW	70% NONE, 20% RARE, 10% SOME
MED	50% NONE, 30% RARE, 20% SOME
HIGH	30% NONE, 30% RARE, 40% SOME

Step 5: Optional Elevation

Zero or one non-emphasized symptom may be elevated (adds unpredictability).

This ensures template-relevant symptoms are more prominent while maintaining overall severity consistency.

6.3.1 Probability-Weighted Coherence

The system uses **weighted probability sampling** throughout profile generation to ensure coherence without rigidity. Rather than deterministically assigning traits, probability weights bias sampling toward psychologically coherent combinations while preserving the possibility of unusual configurations.

Example: A neurotic template tends toward OFTEN-severity depressed mood (80% weight for HIGH intensity), but 20% may yield SOME. This produces variation while maintaining the template's character.

Example: Age-work correlations use weighted pools so students usually fall in 16-29 age ranges but can occasionally be older.

This probabilistic approach creates *diversity without incoherence*. Each generated patient is plausible, and no two are identical. Real populations contain both typical presentations and edge cases.

6.4 Age and Life Context

Age is correlated with work role using **weighted probability sampling**:

Role	Age Weighting
------	---------------

Role	Age Weighting
Students	Heavily weighted toward 16-29
Employed	Spread across working ages (20-59)
Caregiving role	Weighted toward older ages (35-80)
Between roles	Broad distribution

Context domains (0-2 selected) provide stressor themes:

- Work/role strain
- Relationships strain
- Health concern
- Self-worth/identity strain
- Major life transition
- Grief/bereavement

6.5 Background Writer

An LLM generates a rich life background tailored to each patient profile.

Input: Basic profile (age, personality, symptoms, context domains)

Output: Structured life context including:

- Name, age range, pronouns
- Core roles (job seeker, caregiver, student, etc.)
- Key relationships (strained sibling relationship, distant coworkers)
- Core stressor summary
- Life facets (5-8 specific details about goals, coping, history, constraints)

The background writer ensures facets are **age-appropriate** and **severity-appropriate**. Trauma content is gated: mild cases get fewer adversity facets; severe cases may have more.

7. Doctor Configuration

7.1 Doctor Personas

Eight distinct interviewing styles create varied clinical interactions:

Persona	Style Description
warm_validating	Supportive, normalizing, comfortable with pauses
neutral_efficient	Professional, minimal small talk, direct transitions
gentle_brisk	Kind but moves conversation along
matter_of_fact_kind	Straightforward, does not sugarcoat
trauma_informed_slow	Patient, checks comfort, acknowledges difficulty
structured_psychological	Explains why questions matter
time_pressed_clinical	Concise, focused, redirects when needed
dismissive_rushed	Distracted, may minimize concerns

7.2 Microstyle Variation

Each session also samples **per-session variations** on five dimensions:

Dimension	Options
Warmth	low, med, high
Directness	low, med, high
Pacing	slow, med, brisk
Humor	none, light, dry
Animation	reserved, moderate, animated

This means two "warm_validating" doctors may still feel different from each other.

7.3 Emergent Interactions

Doctor-patient combinations create **emergent dynamics**:

Combination	Resulting Dynamic
Warm doctor + guarded patient	Gradual trust building
Dismissive doctor + anxious patient	Increased guardedness
Brisk doctor + verbose patient	Natural tension

These dynamics are **not scripted**. They emerge from the agent interactions.

8. Conversation Orchestration

8.1 Turn Budget System

Each conversation has a budget to ensure all 9 DSM symptoms are covered:

```
target_turns_per_symptom = pacing_target (1.5 to 2.5)
base_budget = target × 9 symptoms
max_turns = base_budget + 5 buffer turns
```

Slower-paced personas get more time.

8.2 The Disclosure Gradient

Patients do not reveal everything at once. **Disclosure evolves** over the conversation.

Initial State (based on trust and verbosity):

- Guarded OR terse → MINIMIZE
- Open AND moderate/detailed → OPEN
- Otherwise → PARTIAL

Evolution:

- FOLLOW_UP from a supportive doctor can increase disclosure
- RAPPORT can gently unlock more sharing
- Guarded patients stay at MINIMIZE longer

Disclosure Stages:

- MINIMIZE: Downplay, deny, underreport symptoms
- PARTIAL: Acknowledge some things, hold back details
- OPEN: Describe symptoms and impact more fully

8.3 Doctor Action Types

The Doctor Manager chooses one of three actions each turn:

Action	Purpose	Example
DSM	Ask about a specific symptom from the checklist	"Have you been feeling down or hopeless?"
FOLLOW_UP	Explore what the patient just said	"You mentioned trouble sleeping. Can you tell me more about that?"

Action	Purpose	Example
RAPPORT	Connect as people, build trust	"That sounds really difficult. How are you managing day to day?"

9. Example Dialogues

The following excerpts demonstrate the **diversity of conversations** the system generates. Each example illustrates different patient profiles interacting with different doctor personas.

9.1 Guarded Patient with Trauma-Informed Doctor

Attribute	Value
Patient Profile	NEUROTICISM_HIGH
Voice Style	terse / guarded / intense
Life Context	35-39 year old caregiver between jobs
Doctor Persona	trauma_informed_slow

This excerpt shows a guarded patient gradually revealing information:

Doctor: Hello. I want you to know this is a safe space. We can take our time today. How are you doing?

Patient: I'm managing, I guess. Just... a bit worn out sometimes. Nothing too drastic.

Doctor: When you say you feel worn out, can you tell me a bit more about that? How often does it happen?

Patient: It's pretty regular, honestly. Most days, I just feel low on energy, like I'm running on empty. Hard to say if anything specific sets it off, just seems to build up over time. Maybe I'm not handling things well enough.

Doctor: It's understandable to feel frustrated with yourself, but I wonder—have you been feeling a lot of guilt or like you're not good enough lately, in a way that's been hard to shake?

Patient: Yeah, I do feel guilty sometimes, like I'm not doing enough. It's hard to let it go—it just lingers. Feels like I'm the problem more than anything else.

Later, discussing suicidal thoughts (note the guarded response):

Doctor: I know this is a hard topic, but sometimes when people feel this overwhelmed and tired, they can have thoughts about death or suicide. Have you had any thoughts like that?

Patient: Sometimes the thoughts pop up, but I don't dwell on them. I try to keep them at bay.

Doctor: How often do they come up, and what helps you push them aside when they do?

Patient: They come up now and then, not all the time. I just try not to let them stick around. It's something I handle on my own.

9.2 Dismissive Doctor with People-Pleasing Patient

Attribute	Value
Patient Profile	AGREEABLENESS_HIGH
Voice Style	moderate / guarded / flat
Life Context	40-49 year old project manager, grief context
Doctor Persona	dismissive_rushed

This excerpt shows how a rushed doctor gets briefer responses:

Doctor: Right, let's get started. What brings you in today?

Patient: Lately, I've just been feeling a bit off—mostly tired and having some trouble sleeping. It's nothing too severe, but it's been enough to notice.

Doctor: Okay, when you say trouble sleeping, do you mean you can't fall asleep, you wake up during the night, or you're up way too early? And how often is this happening?

Patient: I tend to wake up during the night, though it's not every night. It's enough to be noticeable but not consistent.

Doctor: You said concentration is a bit off sometimes—how often do you find it hard to focus or make decisions?

Patient: It happens occasionally, not too often. When it does, it's a bit bothersome, but I'm still able to get my work done.

Observation: Note how the patient minimizes concerns ("nothing too severe," "not too often," "still able to get my work done"), consistent with the people-pleasing, conflict-avoidant profile.

9.3 Open, Articulate Patient with Structured Doctor

Attribute	Value
Patient Profile	CONSCIENTIOUSNESS_HIGH
Voice Style	moderate / intense / open
Life Context	50-59 year old tech manager, suicidal ideation (SOME)

Attribute	Value
Doctor Persona	structured_psychological

This patient is highly articulate, openly sharing detailed information:

Doctor: Hello. Thanks for coming in today. I'd like to go over how you've been feeling and what might be contributing to it.

Patient: Hello. I've been struggling with a persistent low mood that seems to hang over me most days, despite my best efforts to keep some structure in my routine. It's frustrating because, logically, I know that maintaining order should help me feel more in control, yet I often wake up already drained, and that fatigue lingers throughout the day...

My appetite has shifted noticeably too—sometimes I'm almost indifferent to food altogether, which has led to some unwanted weight loss... Sleep—the one thing I hoped might offer relief—has become a battlefield. Nights are restless, fragmented, and often haunted by a mind too busy to rest.

Observation: Note the dramatic contrast with the terse patient in Example 9.1: this patient volunteers detailed, analytical descriptions without prompting.

9.4 Introspective Patient with Metaphorical Expression

Attribute	Value
Patient Profile	OPENNESS_HIGH
Voice Style	moderate / intense / guarded
Life Context	30-34 year old in career transition

This patient uses metaphorical language consistent with the OPENNESS_HIGH philosophical style:

Patient: This shift in my mood has a kind of dulling effect, like colors are less vibrant. It's noticeable but not completely overwhelming, just enough to make things feel a bit muted.

Doctor: Can you tell me more about what kinds of thoughts or feelings tend to pull your motivation down?

Patient: On tougher days, the dips in motivation often feel tied to a sense of futility or a lingering heaviness that hangs around emotionally. There are deeper thoughts beneath that I'm not quite ready to unpack fully yet, but they contribute to that pull. It's a slow undercurrent rather than a sudden wave.

9.5 Patient Discussing Specific Life Context

Attribute	Value

Attribute	Value
Patient Profile	CONSCIENTIOUSNESS_LOW
Voice Style	moderate / intense / open
Life Context	20-24 year old graduate student, part-time barista, recently moved
Doctor Persona	trauma_informed_slow

This excerpt shows a patient referencing specific life circumstances:

Doctor: Can you tell me a bit more about when these tired and restless feelings first started? How do they get in the way of your day-to-day tasks, like studying or working at the café?

Patient: It's been going on for a few months now. At first, I thought maybe it was just a temporary thing, like being overworked or stressed, but it's stuck around longer than I expected. The tiredness and restlessness together really make concentrating a struggle—trying to focus on reading or writing feels draining, like I'm using up energy just to stay present. When I'm at the café, it's hard to slow down or feel settled between tasks, which makes the day more exhausting than it should be.

Later, discussing the impact on daily routines:

Patient: These fluctuations make it really hard to stick to a daily routine. Some days, I have the energy to push through, even though I feel restless and uncomfortable, but that effort just wears me down faster. The disconnect between my restless mind and my tired body is frustrating—it feels like my brain wants to go, but my body holds me back. This tension makes it harder to focus or be productive, especially with school and work.

Observation: The background writer generated a life context including that Maya recently moved to a new city, works as a barista, and is a graduate student—all of which naturally inform her responses without feeling scripted.

10. Output and Data Structure

10.1 What Gets Saved

Each session produces **three output files**:

File	Content	Primary Use
<code>transcript_AGENT_xxx.json</code>	Dialogue + metadata + ground truth	ML training data
<code>session_AGENT_xxx_raw.json</code>	Full details + all intermediate outputs	Debugging & analysis
<code>prompttrace_AGENT_xxx.json</code>	All prompts sent to all agents	Prompt engineering

10.2 Transcript Structure

The primary output file contains all information needed for downstream ML tasks:

```
{
  "agent_id": "AGENT_abc123...",
  "template_id": "NEUROTICISM_HIGH",
  "persona_id": "warm_validating",

  "personality": {
    "VOICE_STYLE": {
      "verbosity": "moderate",
      "trust": "guarded",
      "expressiveness": "intense"
    },
    "MODIFIERS": ["worry-prone", "self-blaming"],
    "AGE_RANGE": "30-34"
  },

  "depression_profile_ground_truth": {
    "Depressed mood": "OFTEN",
    "Sleep disturbances": "SOME",
    "Fatigue": "OFTEN",
    "Concentration": "RARE"
  },

  "conversation": [
    {"speaker": "doctor", "text": "Hello! How have you been?"},
    {"speaker": "patient", "text": "I've been okay, I guess."}
  ],

  "doctor_manager_decisions": [...],
  "patient_manager_decisions": [...],
  "final_disclosure_state": "PARTIAL"
}
```

10.3 Agent ID

Each unique configuration gets a **deterministic ID** from hashing all parameters:

- Same parameters → same ID (enables reproducibility)
- Different parameters → different ID (enables duplicate detection)

This allows tracking which configurations have been generated and ensures dataset diversity.

11. Current Limitations

11.1 Cost and Speed

Metric	Value
Cost per conversation	\$0.01–\$0.20 (depending on model)
Generation time	1–3 minutes per session

11.2 Occasional Generation Artifacts

Under certain conditions (weak models, edge case configurations), conversations may have minor inconsistencies. Mitigation strategies:

- Multiple **fallback mechanisms** prevent complete failures
- **Volume-based filtering** removes noise from large datasets
- **Quality scoring** (future work) could flag problematic outputs

11.3 Text-Only Format

Feature	Status
Emotional tags in dialogue	<i>Removed</i>
Stage directions	<i>Removed</i>
Emotional metadata in JSON	Preserved

The current text-only format was chosen for compatibility with downstream ML tasks. All emotional metadata (tone tags, manager instructions) is saved in transcripts, preserving the option to use this information later.

12. Future Work

12.1 Near-Term Development

Initiative	Description
More model options	Support for additional LLM providers and local models
Expanded diversity	Additional doctor personas, Big Five templates, voice dimensions

Initiative	Description
Prompt tuning	Current prompts have not been exhaustively optimized

12.2 Longer-Term Research Direction

A promising direction involves adding a **third processing layer** to separate speech content from speech form.

Current State: The patient agent handles both:

- *What* the patient communicates (symptoms, feelings, facts)
- *How* the patient expresses it (word choice, phrasing, language patterns)

Proposed Separation:

Layer	Responsibility
Content layer	Generate what the patient wants to convey (current system)
Style layer	Transform the content into a specific language form

This separation could enable simulation of linguistic features that current single-layer approaches handle poorly:

- Non-native English speakers
- Regional dialects
- Speech patterns associated with various conditions

By having one layer responsible for content accuracy and a separate layer for surface presentation, we could achieve demographic diversity without the stereotyping problems that plague direct prompting approaches.

This approach mirrors findings from *SpeechDialogueFactory* that separating content generation from paralinguistic features improves quality in multi-modal synthesis.

13. Applications

13.1 Potential Uses

The generated dataset can serve multiple purposes:

Application	Description
Training ML models	Primary intended use—labeled conversation data for models that analyze or participate in mental health dialogues
Clinical training	Synthetic patients for practicing screening conversations without standardized patient actors
Research	Studying how different interviewing styles affect patient disclosure

13.2 Adaptability

The system is designed to be **easily modified** for different use cases:

Use Case	Modification
Questionnaire-based screening	Simplify doctor manager to follow fixed question sequence; patient diversity unchanged
Voice synthesis	Emotional metadata (tone tags, manager guidance) can drive TTS systems with varied prosody
Other conditions	Replace DSM depression criteria with anxiety, PTSD, or other screening items

The multi-agent architecture **generalizes beyond depression**—the patient diversity framework remains applicable to any clinical dialogue domain.

14. Usage Guide & Extending the System

14.1 Setup

Installation:

```
pip install openai python-dotenv
```

API Key Configuration:

```
# Option 1: Environment variable  
export OPENAI_API_KEY=your-key-here  
  
# Option 2: Create .env file  
echo "OPENAI_API_KEY=your-key-here" > .env
```

14.2 Basic Usage

```
python -m synthetic_datagen.cli --num-sessions 5
```

14.3 Forcing Configurations

For testing specific patient/doctor combinations:

Flag	Purpose	Example
--forced-template	Force Big Five template	NEUROTICISM_HIGH
--forced-trust	Force trust level	guarded
--forced-verbosity	Force verbosity	terse
--forced-age	Force age range	"70-80"
--forced-density	Force symptom density	ULTRA_LOW
--forced-persona	Force doctor persona	dismissive_rushed

Example:

```
python -m synthetic_datagen.cli \  
    --forced-template NEUROTICISM_HIGH \  
    --forced-trust guarded \  
    --forced-persona trauma_informed_slow
```

14.4 Test Modes

Zero API cost testing:

Mode	Command	Purpose
Profile test	--test-profile	Test profile generation only
Manager test	--test-patient-manager	Test patient manager with mock data

```
# Test profile generation (no API calls)
python -m synthetic_datagen.cli --test-profile

# Test patient manager logic
python -m synthetic_datagen.cli --test-patient-manager
```

14.5 Adding Doctor Personas

Location: [synthetic_datagen/prompts/doctor_personas.py](#)

Add to the `DOCTOR_PERSONAS` dictionary:

```
{
    "id": "your_persona_name",
    "first_greeting": "Opening line here",
    "system_prompt": """
        Persona description including:
        - Communication style
        - Typical behaviors
        - How they handle difficult topics
    """,
    "pacing_target": 2.0 # turns per symptom
}
```

14.6 Adding Big Five Templates

Location: [synthetic_datagen/data/templates.py](#)

Add to the `BIG5_DEP_TEMPLATES` dictionary:

```
"YOUR_TEMPLATE": {
    "affective": "Emotional style description",
    "cognitive": "Thinking pattern description",
    "somatic": "Physical manifestations",
```

```
"emphasized_symptoms": [  
    "Symptom1",  
    "Symptom2",  
    "Symptom3"  
,  
    "modifiers": [  
        "trait-modifier-1",  
        "trait-modifier-2"  
    ]  
}
```

Note: Emphasized symptoms should be valid DSM-5 symptom keys from [DSM5_ITEMS](#).

15. References

Patient Simulation Systems

Reference	Venue
PatientSim: A Persona-Driven Simulator for Realistic Doctor-Patient Interactions	arXiv:2505.17818
PATIENT-Ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals	arXiv:2405.19660
Depression Diagnosis Dialogue Simulation (D4)	EMNLP 2022
MDD-5k: Multi-round Doctor-Patient Dialogues for Major Depressive Disorder Diagnosis	arXiv:2408.12142
PSYCHE: A Multi-Faceted Patient Simulation Framework for Psychiatric Assessment	—

Communication Style and Persona Research

Reference	Venue
Beyond the Script: Testing LLMs for Authentic Patient Communication Styles in Healthcare	arXiv
Persona-Augmented Benchmarking: Evaluating LLMs Across Diverse Writing Styles	EMNLP 2025
Scaling Synthetic Data Creation with 1,000,000,000 Personas (Persona Hub)	arXiv:2406.20094

Sociodemographic Prompting Studies

Reference	Venue
Beck et al. (2023): Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting	arXiv:2309.07034
Sun et al. (2025): Sociodemographic Prompting is Not Yet an Effective Approach	NAACL 2025

Dialect and Bias Research

Reference	Venue
Finch et al. (2025): Finding A Voice: Evaluating African American Dialect Generation for Chatbot Technology	ACL 2025
Hofmann et al. (2024): AI generates covertly racist decisions about people based on their dialect	Nature 633, 147–154

Appendices

Appendix A: DSM-5 Depression Symptoms

The **9 symptoms** from DSM-5 criteria for Major Depressive Episode:

#	Symptom	Clinical Description
1	Depressed mood	Most of the day, nearly every day
2	Loss of interest or pleasure	Markedly diminished interest in activities
3	Weight/appetite changes	Significant weight loss/gain or appetite change
4	Sleep disturbances	Insomnia or hypersomnia
5	Psychomotor changes	Agitation or retardation observable by others
6	Fatigue	Loss of energy nearly every day
7	Worthlessness/guilt	Excessive or inappropriate guilt (may be delusional)
8	Concentration problems	Difficulty thinking or indecisiveness
9	Suicidal ideation	Recurrent thoughts of death (not just fear of dying)

Diagnostic threshold: ≥5 symptoms present during same 2-week period, representing change from previous functioning. Must include either (1) depressed mood or (2) loss of interest/pleasure.

Appendix B: Complete Voice Style Reference

Verbosity

Option	Behavioral Effect
terse	Short responses, minimal elaboration
moderate	Average length, some context
detailed	Longer responses, provides background

Expressiveness

Option	Behavioral Effect
flat	Limited emotional range, monotone
balanced	Normal emotional expression
intense	Strong emotional display

Trust

Option	Behavioral Effect
guarded	Suspicious, careful disclosure
neutral	Normal wariness
open	Willing to share readily

Intellect

Option	Behavioral Effect
low-functioning	Simple vocabulary, concrete thinking
moderate-functioning	Average cognitive presentation
high-functioning	Complex vocabulary, abstract thinking

Humor

Option	Behavioral Effect
none	No humor used
occasional	Light humor sometimes
frequent	Often uses humor (often as coping)

Appendix C: Episode Density Details

Density	Target Profile	Non-Emphasized Symptom Weights
ULTRA_LOW	0-2 symptoms total	N/A (special mode)
LOW	Light symptoms	70% NONE, 20% RARE, 10% SOME
MED	Moderate symptoms	50% NONE, 30% RARE, 20% SOME
HIGH	Severe symptoms	30% NONE, 30% RARE, 40% SOME

Note: Emphasized symptoms (determined by Big Five template) use **intensity-based sampling** regardless of episode density.

Document generated for the Depression Detection Research Project

November 2025