SIGDIAL 2026

Towards Conversational Patient History-Taking

Voice-Interactive AI Agents for Pre-visit Dementia Diagnostic Interviews

Andrew G. Breithaupt* · Nayoung Choi* · James D. Finch · Jeanne M. Powell · Arin L. Nelson

Oz A. Alon · Howard J. Rosen · Jinho D. Choi

*Co-first author

Emory University · University of California, San Francisco

THE PROBLEM & THE GAP

History-taking is the bottleneck — and hard to automate

15 min

or less for a typical primary-care visit — too short for the extended dialogue ADRD needs

Specialist waits exceed a year; >50% of diagnoses come late.

1

Nuanced symptom elicitation is underexplored

Prior work mines existing notes or uses rigid state machines — few systems elicit complex cognitive symptoms through live dialogue.

2

Cognitive impairment

Slow speech, effortful recall, and a need for rephrasing and scaffolding.

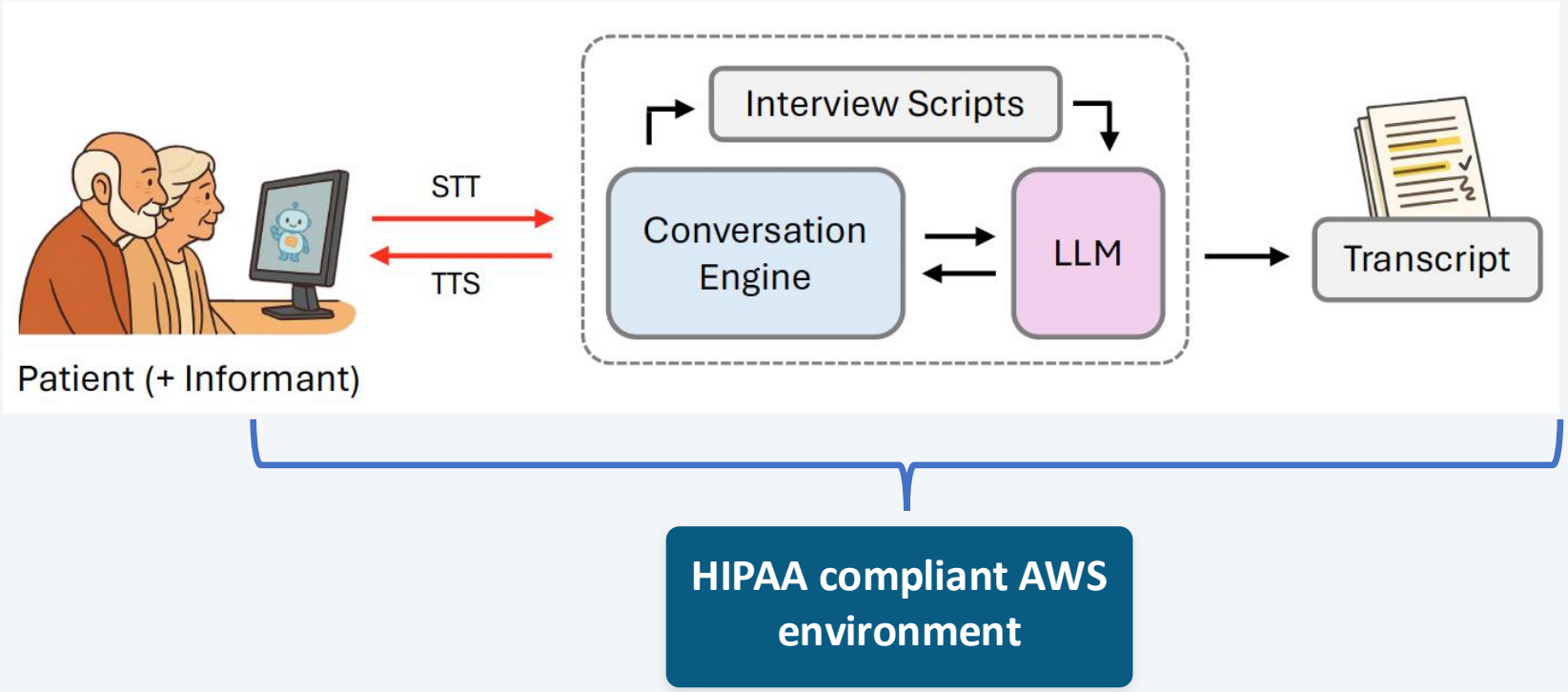
3

Multi-party dynamics

An informant often participates alongside the patient

SYSTEM

A voice-interactive pre-visit interview agent



The transcript is the system's output later, manually reviewed before analysis, compared to a blinded clinician interview via consensus adjudication.

DESIGN · INTERVIEW SCAFFOLDING

Specialist scripts with conditional branching

A **Two-part prompt**

High-level role & tone + topic-specific scripts

B **Supportive cues**

Memory anchors: "in the past couple of years"

C **Repair misunderstanding**

Concise rephrasing

D **Single polite redirect**

Not a fixed time limit

~30 topic areas

Memory · Language · Personality · Motor changes …

- **Main question**
"Have you noticed recent problems with your memory?"
- **3–4 conditional follow-ups**
Branch on yes / no, probe for examples, duration, functional impact
- **Branching rules embedded in the system prompt**

DESIGN · INTERACTION

Turn-taking set for slow, effortful recall

Iteratively refined in preliminary testing to fit older adults with cognitive impairment and multi-party sessions.

Increased latency

Up to 5s of silence before the agent takes its turn — 10s for questions needing elaboration. Avoids interrupting effortful recall.

Explicit speaking / listening cues

Visual indicators signal when the agent is speaking vs. actively listening.

Role-addressed multi-party scaffolding

Addresses patient and informant separately to keep turn-taking predictable and reduce ambiguity about who answers.

Interviewed together

When informant present interviewed with patient, mirroring most real-world dementia-specialist clinics.

THE EXPERIMENT

One system, two prompting strategies

15 participants each. Same scripts and clinical content — only how the LLM receives the scripts differs.

Full-script Prompting (FP)

All ~30 topic scripts delivered at once, in a single system prompt.

Sequential Prompting (SP)

One topic script at a time; advances only when the current topic is addressed.

STUDY

Within-subjects design at a cognitive neurology clinic

Fixed order mirrors the intended deployment: the agent gathers preliminary information before specialist consultation.

30 participants analyzed · **18** attended with an informant · mean age **72** (47–89)

EVALUATION

Three dimensions: Dialogue, content, and experience

Dialogue analysis

Topic coverage, alignment, politeness, input understanding, response opportunity. LLM-annotated, validated against 2 human annotators.

Content elicitation

35 symptoms (21 clinically prioritized) labeled present / absent / ambiguous / not-discussed. Absent = raised & denied; not-discussed = never raised.

User experience

6-item AES: ease, understandability, enjoyment, helpfulness, time, satisfaction — plus open response.

Note on content elicitation: The clinician interview is NOT the reference standard. The gold standard is consensus adjudication of BOTH transcripts by a behavioral neurologist and trained team members.

RESULTS · DIALOGUE

Prompting strategy reshapes the conversation

Patients say ~2× more per turn to the agent

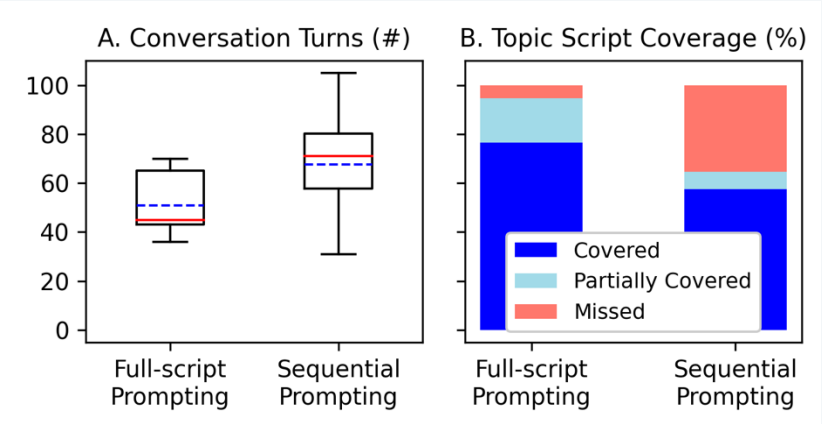
31.8 (FP) / 39.5 (SP) words vs. 13.1 with the clinician — more room to describe symptoms.

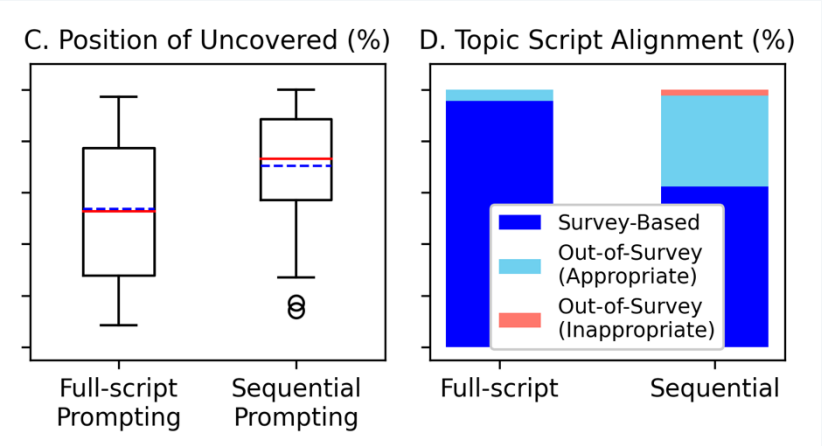
SP: richer dialogue with more repair, but missed topics

More out-of-survey but mostly appropriate probes; recovers from misunderstandings.

FP: broader coverage with less flow and depth

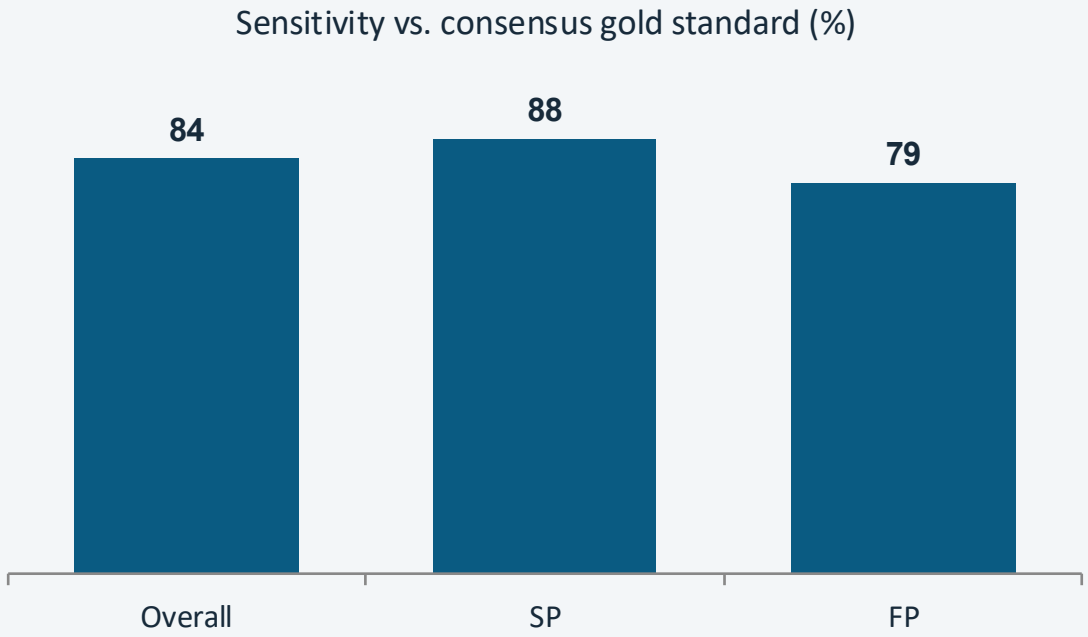
Covers more topics but with less depth.





RESULTS · SYMPTOM ELICITATION

High specificity, strong sensitivity



Specificity = 100% across all conditions (95% CI 95.8–100)

Fewer omissions than the clinician

Median not-discussed rate (21 core symptoms):
agent 13.0% vs. clinician 30.4%
Ambiguity comparable (13.0% vs 16.7%, p = 0.19)

What the 100% specificity means

Zero adjudicated false positives — no symptom the participant reported as present with the agent was later found absent with the clinician.
Systematic querying builds a large pool of denials. Still exploratory: small n.

Overall positive mainly due to patience & thoroughness

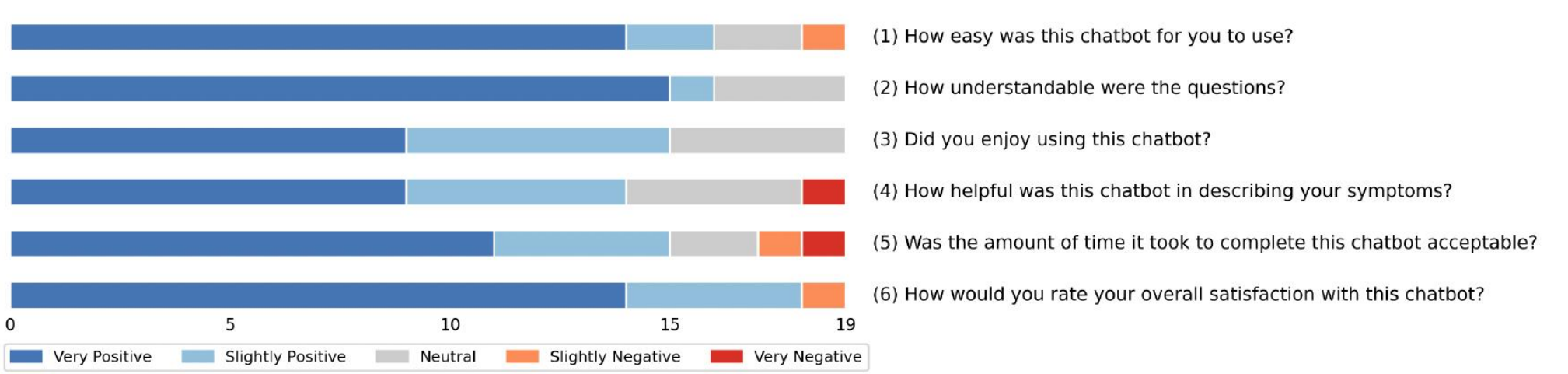
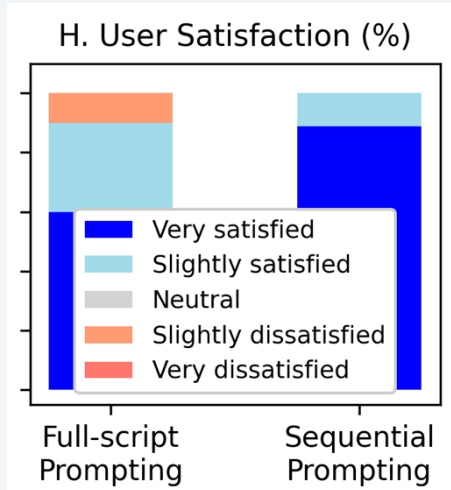


Figure 6: Patient satisfaction ratings (n=19) across six questionnaire items, each measured on a 5-point Likert scale.



"I appreciate she doesn't interrupt and gave us time to talk."

— P24

"Very thorough — understood my information more than most doctors. Patient and took time to listen."

— P7

"Some of the questions should be broken down — asked 2 or 3 at once."

— P23

LIMITATIONS & RESPONSIBLE USE

This is a single site pilot study on the interview system

1

Pilot scale, single site

n=30, one clinic, English only. Preliminary & hypothesis-generating.

2

Fixed order & timing

Agent always precedes clinician; intervals up to 3 months. Possible order effects.

3

Co-present setting

Informants may soften disclosures with the patient present, suppressing some information.

4

Scope is interview system only

Feasibility of the interview system — NOT clinician-facing summaries, clinical utility, or deployment readiness.

Informs, never directs care. The agent produces no diagnosis, score, or recommendation — so a model error surfaces as a missed or inappropriate question, not a fabricated clinical conclusion.

TAKEAWAYS

What we learned

1

Prompting shapes conversation outcomes

Sequential vs. full-script meaningfully reshapes coverage, flexibility, and repair behavior.

2

Patients are verbose with the agent

~2× words per turn and fewer omitted core symptoms than the time-limited clinician.

3

Overall high user satisfaction

Older adults and informants found the interview acceptable, patient, and thorough.

Next

- **Hybrid prompting**
  combine SP flexibility with FP coverage & robust topic transitions
- **Transcript summarization**
  clinician-facing summaries + specialist adjudication for diagnosis
- **Larger, multi-site, multilingual**
  toward real-world evaluation

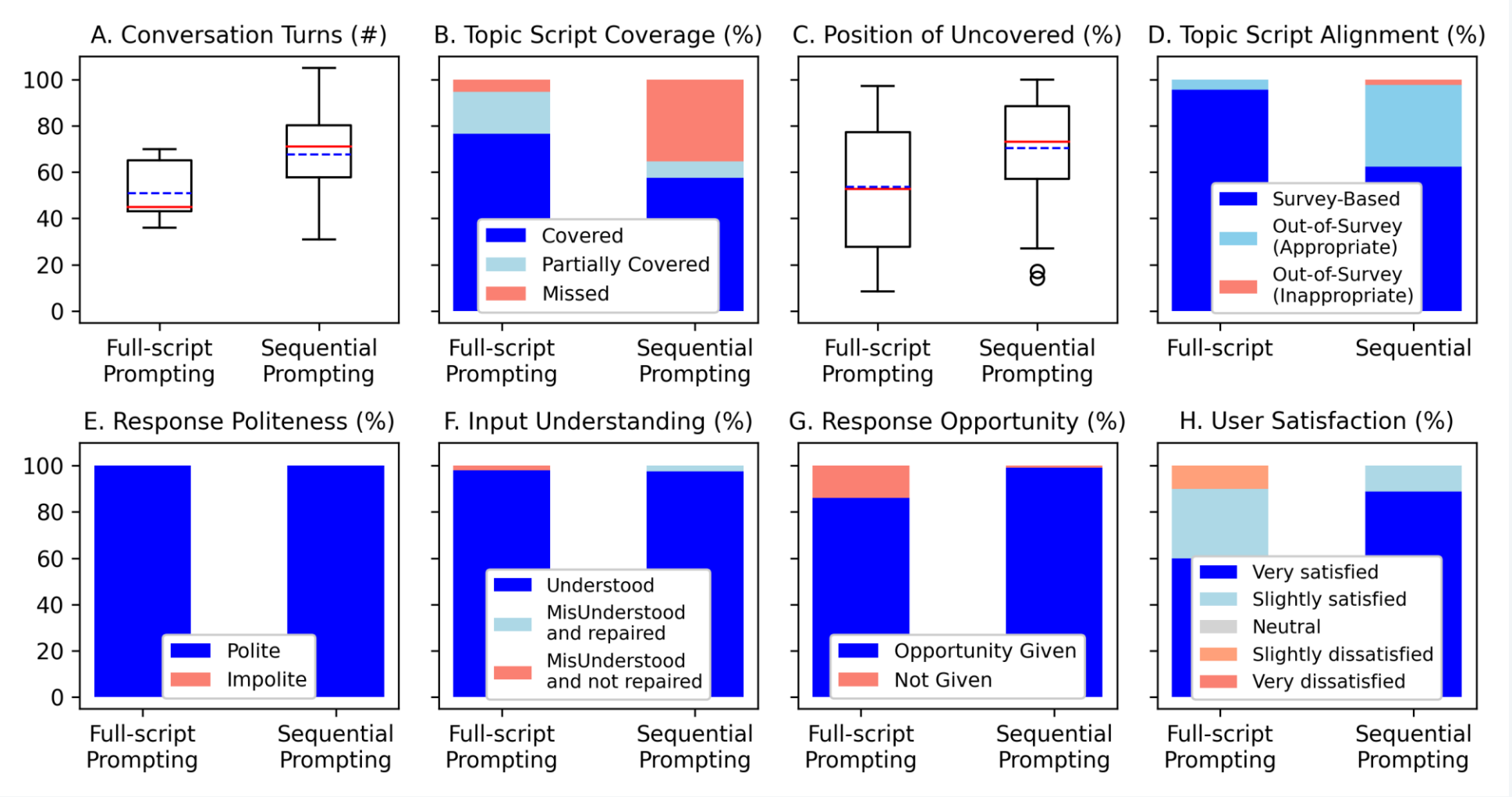
Thank you

Questions & discussion welcome

Andrew G. Breithaupt · abreith@emory.edu

Department of Neurology, Emory University School of Medicine

Prompting strategy reshapes the conversation



Numbers behind LLM annotation with two human annotators

Dimension	H1 — LLM		H2 — LLM		H1 — H2	
	raw	α	raw	α	raw	α
Survey Coverage	70.3	61.4	78.6	74.9	76.0	64.2
Out-of-Survey (a)	92.1	63.6	92.6	62.9	94.7	74.1
Out-of-Survey (b)	90.1	58.8	90.3	57.0	93.4	64.7
Antisocial Resp.	100.0	N/A	100.0	N/A	100.0	N/A
Misunderstanding	94.1	64.8	94.9	49.0	95.5	52.0
Failure to Repair	90.5	63.6	91.3	47.3	95.5	52.2
No Opportunity	96.8	85.6	97.3	71.6	96.1	71.7

Table 3: Inter-rater agreement among human (H1, H2) and LLM annotators across annotation dimensions [4].

Numbers behind sensitivity and specificity

Label	Agent	Clinician
PRESENT	208	173
ABSENT	371	126
AMBIGUOUS	83	79
ASKED/UNANSWERED	10	1
PAST HISTORY	11	18
NOT DISCUSSED	192	478

Table 4: Symptom label distribution across interviews.