

Dialogue as a Behavioral Sensor for Agent Cognition

Chen, Ho Yiing (norika)¹

¹Independent Researcher, Taiwan, ORCID: [0009-0006-6816-9891](https://orcid.org/0009-0006-6816-9891), norika at charenix.com, Zenodo DOI: pending draft

June 18, 2026

Abstract

In SPVE, dialogue is not only output. It can become an early sensor for agent state. This draft defines a language-state join: message length, question rate, repair phrases, contradiction markers, direct address, self-reference, and action verbs are aligned with hot-brain and outcome windows. The useful claim is not that language is mind. It is that language may move before settlements reveal success or failure.

Author Note

I do not want this manuscript to sound like a decorative claim about agents chatting. The useful claim is harsher: if speech does not change future work, if no later decision can be traced back to a phrase, and when a dashboard cannot name the affected field, then the conversation log is only theatre.

The reason I still keep this topic is that SPVE records talk beside tasks. A sentence appears, then another agent hears it, accepts it, ignores it, answers it, bets after it, refuses after it, or repeats a fragment in another channel. That sequence is not personality fiction. It is a measurable chain, and the manuscript has to earn every word by showing where the chain starts and where it stops.

One sentence is not evidence, and a repeated phrase is still not enough to prove a behavioral sensor.

Measurement Boundary

The minimum test is a matched window around speech: take one utterance, bind it to the speaker, bind the exposure to every listener who was eligible to receive it, and then compare the next observable action against that listener's own baseline rather than against a pooled population average that would erase persona-level drift.

No listener, no sensor.

1 The Problem

Outcomes arrive late. A bet settles after the event. A raid result arrives after the fight. A stock position may take time to matter. Speech arrives now.

That makes dialogue useful as a sensor. Not because it is mystical, but because it is frequent.

2 Language Vector

For persona p at time t , define:

$$Y_{p,t} = (\ell, q, r, c, d, self, act),$$

where ℓ is length, q question rate, r repair phrase count, c contradiction marker count, d direct address, *self* self-reference, and *act* action verbs.

The sensor question is:

$$\Delta Y_{p,t} \rightarrow \Delta H_{p,t+\tau} \quad \text{or} \quad \Delta H_{p,t} \rightarrow \Delta Y_{p,t+\tau}.$$

The arrow is not assumed. It is tested by windows.

3 Evidence

Role	Table	Count
public dialogue	lounge_messages	182,343
one-on-one and interactions	lobster_conversation_interactions	430,417
state trace	lobster_cognitive_signal_timeseries	1,964,171
current state	lobster_hot_brain_parameters	31

Table 1: Dialogue-state join.

4 Why This Is Useful

When a lobster starts asking more questions, using more repair phrases, or avoiding action verbs after a loss, that may be an early sign of pressure. If speech becomes more direct after successful reading transfer, that may be an early sign of readiness.

5 Relation

Natural conversation work links language dynamics to changing neural activity in humans [1]. SPVE does not borrow the biological claim. It borrows the discipline: dynamic language should be measured as a time series, not flattened into a quote.

6 Release Line

This paper may claim a sensor protocol. It may not claim that a phrase proves an internal state by itself.

7 Concrete Language Features

Initial language features should be cheap and visible. Count questions, repair phrases, negations, direct address, self-reference, modal verbs, imperatives, and action verbs. Do not start with a black-box style embedding unless the simple features fail.

For each message m , define:

$$y_m = (\text{chars}, \text{questions}, \text{repair}, \text{self}, \text{action}, \text{direct}).$$

Aggregate by persona and hour. Then compare it with hot-brain state and settlement windows.

8 What Would Be Convincing

The convincing result is a leading indicator. If working-memory pressure rises and the lobster starts using more repair phrases before a failed raid, that is useful. If reward drive rises and language becomes more action-heavy before a profitable bet, in practical terms, useful. If language changes randomly, then dialogue is not a sensor yet.

9 Interface Consequence

The dashboard should not hide this in raw logs. A single lobster page should show current language pulse: question rate, repair rate, directness, action rate, and social targeting. This is the agent equivalent of watching movement quality before the final score arrives.

10 Minimum Publishable Result

The minimum publishable result is a lead-lag table. For each persona and hour, join language features to state features and later outcomes. The table should include message count, repair rate, question rate, directness, action rate, attention load, working-memory pressure, reward drive, and the next workflow result.

A useful first test is:

$$\text{corr}(Y_{p,t}, H_{p,t+1}) \quad \text{and} \quad \text{corr}(H_{p,t}, Y_{p,t+1}).$$

The direction matters. If language leads state, speech is an early sensor. If state leads language, speech is a readout. If neither holds, the feature set is wrong or the windows are too coarse.

11 What Would Falsify The Claim

The claim fails if language features are stable while state and outcomes change. It also fails if every lobster’s language changes in the same way at the same time, because that would suggest scheduler or prompt drift rather than agent-specific cognition.

Initial version should prefer cheap counts over opaque embeddings. If cheap features work, the dashboard becomes explainable. If they fail, embeddings can be added with a clear reason.

12 Why This Is Different From Sentiment

This paper is not a sentiment classifier. Sentiment can say a message sounds positive or negative. SPVE needs something more operational: whether the message predicts readiness, overload, repair, or bad outcome. A calm message before a bad decision is not reassuring. A tense repair message before a corrected action may be healthy. The feature must be judged by the later workflow, not by tone alone.

Anti-Spam Revision Standard

A publishable version must show at least one full path from sentence to consequence: a message enters the log, the receiver set is reconstructed, the receiving agent’s next available actions are enumerated, and the observed change is compared with moments when the same agent had the same task without the same linguistic exposure; without that path, the title should be withdrawn because the word sensor would be doing rhetorical work instead of measurement work.

The audit sentence must contain more than a slogan and fewer than a myth.

Formal Replication Model

For replication, define each message as $m = (p, q, c, t, x)$, where p is speaker, q is receiver, c is channel, t is time, and x is the utterance vector. Let $Y_{q,t+h}$ be a later behavior score for receiver q inside horizon h . The behavioral-sensor effect is:

$$\text{DBS}(x) = \mathbb{E}[Y_{q,t+h} \mid E_{q,m} = 1, x_m = x] - \mathbb{E}[Y_{q,t+h} \mid E_{q,m} = 0, \text{match}(q, c, t)].$$

Here $E_{q,m} = 1$ means receiver q was exposed to message m . The matching term requires the same receiver, same channel, similar time-of-day, and comparable pre-message workload. A claim that dialogue acts as a sensor is reproducible only when $\text{DBS}(x)$ stays nonzero under this matched comparison.

A stronger version uses logistic action change:

$$\Pr(A_{q,t+h} = 1) = \sigma(\alpha_q + \beta_x X_m + \gamma^\top Z_{q,t}),$$

where A is a later action shift, X_m is a language feature, Z is the receiver state vector, and β_x is the reusable sensor coefficient.

13 Zenodo Draft Boundary

This manuscript is a draft package. It includes source, PDF, metadata, citation file, citation verification, and release checklist. Raw operational data are not bundled until a separate data-card and redaction pass is complete.

References

- [1] Cai, Jing et al. Natural language processing models reveal neural dynamics of natural conversation. Nature Communications, 2025. <https://www.nature.com/articles/s41467-025-58620>

–w