INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 16, Issue 9, (2026) E-ISSN: 2222-6990

A Voice-and-Vision Dialogue System for Observing Emergent Cognition in Learners: Design Rationale and What the Cloud cannot Provide

Naoki Ohshima*

Graduate School of Innovation and Technology Management, Yamaguchi University 2-16-1, Tokiwa-Dai, Ube, Yamaguchi, 755-8611, Japan

*Corresponding Author Email: nohshima@yamaguchi-u.ac.jp

DOI Link: http://dx.doi.org/10.6007/IJARBSS/v16-i9/29073

Published Date: 23 September 2026

Abstract

The moments at which a learner's understanding reorganises itself sit at the heart of teaching practice, yet they are among the hardest things to observe from the outside. This paper proposes reading the internal states of a dialogue model not as a measure of the model, but as an instrument for measuring the person in front of it, inverting surprisal theory from comprehension to production. As a first step, we built a preliminary system coupling a conference speakerphone and a PTZ camera to a cloud speech dialogue API, and characterised it as an observation instrument. The subject was the author alone (self-experimentation). Both the spoken dialogue loop and image-grounded dialogue were established: the noise floor measured −63.4 dBFS against speech at −22 to −28 dBFS, and the model read back a string printed on the subject's shirt, information available only from that frame. Against this, we observed responses describing a visual input in detail when none had been sent, even though the recogniser had correctly registered its absence. A control turns under near-identical conditions asked appropriately for repetition instead, and nothing in the returned output distinguishes the two. The obstruction proved consistent across every stage of the observation chain: the microphone's signal processing exposes no parameters and cannot be disabled, and the API returns neither log probabilities for the learner's speech, nor hidden representations, nor attention weights. In the visual case, we could adjudicate the confabulation against the frame we had sent; a learner's inner state affords no such original. For an instrument, the inability to trace why a response was produced is disqualifying. What this research requires, then, is not better performance but ownership of the instrument.

Keywords: Cognitive Observation, Spoken Dialogue System, Confabulation, Surprisal, Educational Technology

Introduction

The moments when a learner's understanding deepens, when previously separate ideas suddenly connect, are central to educational practice yet remarkably resistant to observation. This paper focuses on adult learners, and the account that follows is framed accordingly. Three properties account for this resistance.

First, such moments are accessible only from within. An adult learner can reflect on their own process, and the moment at which understanding reorganises itself is registered internally; what remains unavailable is any external view of it. Asking afterwards retrieves a reconstruction rather than the moment itself, which makes retrospective self-report an unreliable instrument for locating when it occurred.

Second, they leave no trace. Written text preserves only the settled state that follows revision; the hesitation, the false start, the line of reasoning taken up and abandoned have all been edited out by the time a sentence is finished.

Third, the act of observing alters what is observed. Attaching a sensor, asking a direct question, or simply making the learner aware of being recorded all change the behaviour under study.

This work chooses speech not for convenience but because speech is unrevised. Utterance leaves the speaker at very nearly the same moment as the thought that produced it. Hesitations, the length of a pause, self-repairs, broken word order, all of these survive in audio and vanish from writing. This is precisely why think-aloud protocols have relied on speech (Ericsson & Simon, 1993). Speech also offers a practical advantage: it can continue while the hands are occupied and the eyes are elsewhere, increasing both the frequency and naturalness of what is captured.

A large language model (LLM) maintains, internally, a probability distribution over the next token given its context. Surprisal (negative log-likelihood) and entropy derived from that distribution have conventionally been treated as quantities that evaluate the model's own linguistic competence. This work inverts that direction. We propose treating the model as a fixed measuring device and reading the prediction error arising inside it as an index of the moment when a human speaker departs from the routine.

The idea rests on established psycholinguistic ground. Surprisal theory (Hale, 2001; Levy, 2008) formalises processing difficulty as the difficulty of predicting the next word from context; reading times scale logarithmically with word predictability (Smith & Levy, 2013), and correspondences with the N400 component have been reported (Frank et al., 2015). LLM-derived surprisal accounts for these human measures and is now well established. The novelty lies in the application direction (Table 1).

Table 1
Conventional surprisal research and the present work

	Conventional surprisal research	Present work
Subject	A person reading or listening	A person speaking
What is explained	Processing load in comprehension	Cognitive leaps in production
Experimental form	Controlled stimulus sentences	Free speech
Purpose	Validating models of language processing	Observing learners

Four internal quantities are of interest. Surprisal may indicate departure from the routine; entropy, whether thinking stands at a branch point or is converging; the trajectory of hidden representations, shifts of topic or reframing; and the referents of attention, the reconnection of distant memory. Of these, an utterance that is high in surprisal at a moment when entropy is also high is the most promising composite index: it marks a point at which the path taken was unexpected even though the alternatives were open.

This paper reports the stage preceding that methodology. We built a preliminary system on a cloud API and evaluated its properties as an observation instrument. Our contributions are threefold: establishing a voice-and-vision dialogue loop together with a quantitative characterisation of the recording environment; demonstrating empirically that the obstructions to observation are consistent across every stage of the chain, from the DSP inside the microphone to the cloud LLM itself; and classifying responses that describe non-existent inputs in detail, clarifying what such responses imply for the use of the system as an instrument.

The author conducted the study alone (self-experimentation). Verification with learners requires ethical review and is deferred to the next stage.

This work was carried out as part of a research project on developing a generative-AI-based adaptive learning support environment for secondary education (JSPS KAKENHI 26K05986), specifically as an investigation of the elemental technologies underpinning its observation layer. Designing such an environment requires a way to record what a learner is coming to understand, and how; this paper addresses that observational side.

Literature Review

This work focuses on adult learners. Knowles (1980) distinguished adult education from that of children, arguing that adult learners direct their own learning and draw on accumulated experience as a resource. The distinction matters here for a specific reason. An adult learner can reflect on their own process, so the moment understanding reorganises itself is registered internally even when nothing of it is visible from outside. What is missing is not the learner's awareness but an external record of it, and this work aims to provide that.

The think-aloud protocol is well established as a method for inferring cognitive processes from speech recorded during task performance (Ericsson & Simon, 1993). Chi (1997) set out a systematic procedure for treating such verbal data quantitatively. These methods, however, presuppose manual coding. An analyst transcribes the speech, defines categories, and assigns

segments. The process is laborious, inter-rater agreement becomes a concern, and, most limiting of all, it cannot operate in real time. Our interest in automatically derived internal quantities is motivated by the wish to sidestep this constraint.

The sustained practice of writing about one's own experience promotes reflection and reconstruction (Pennebaker, 1997). Schön (1983) placed the practitioner's reflection, both in and on action, at the centre of professional expertise. Voice journaling removes the burden of writing from that practice. However, recordings that merely accumulate as audio do not advance analysis. The presence of an interlocutor changes the record from monologue into an act that draws a response. This is why the present work builds a dialogue system rather than simply a recorder.

Reeves and Nass (1996) demonstrated across a wide range of experiments that people respond socially to computers as a matter of course. Eliciting that response does not require a humanlike form. Keepon (Kozima et al., 2009) achieved expressions of attention and affect with nothing more than a yellow snowman-like body and four degrees of freedom. As Breazeal (2003) argued, what matters in a social robot is not fidelity of form but the timing and contingency of movement. The PTZ camera used here has two degrees of freedom, pan and tilt, fewer than Keepon, but enough for the single gesture of turning to face something.

Hale (2001) and Levy (2008) formalised processing difficulty in comprehension as the difficulty of predicting the next word from context. Smith and Levy (2013) established the logarithmic relationship with reading time; Frank et al. (2015) showed correspondence with event-related potentials. All of this work addresses comprehension. Applying the same quantities to production, to the speaker's own utterance, read as an index of cognitive leaps, has not, to our knowledge, been pursued systematically.

That generative models produce fluent content contrary to fact is widely documented (Ji et al., 2023; Maynez et al., 2020). Most of this literature concerns factual consistency between generated text and a reference document. The cases reported here arise at a different level. The input itself is fabricated, the model describes in detail a modality that was never transmitted, and frameworks that presuppose comparison against a reference do not capture this. To mark the distinction, we use the term *confabulation* rather than *hallucination* throughout this paper.

Design Rationale

The system is deliberately built as three separable layers: the setting, the voice, and the intelligence (Table 2). This separation lets us assess the constraints of each layer independently. As we show below, obstructions to observation exist in both the setting layer and the intelligence layer, and they differ in kind.

Table 2
The three layers and their implementation

Layer	Responsibility	Implementation here
Setting	Physical conditions of capture and playback	Yamaha YVC-300
Voice	Sustaining the dialogue	Gemini Live API
Intelligence	Obtaining and interpreting internal quantities	Not implemented in this paper

Fitting a learner with a headset or lapel microphone would improve capture quality. However, fitting it announces that recording is about to begin, and speech becomes guarded. We therefore use a conference speakerphone placed on the desk. This is a deliberate trade: we accept degraded capture quality in exchange for a reduced observer effect. To test whether the trade is defensible, we quantitatively characterised the resulting capture.

The system includes a PTZ camera, which serves two roles. As a neck, for expression, pan and tilt convey the direction of attention and something of affect; because the form is not humanlike, a social response can be elicited without raising expectations of full anthropomorphism. As an eye, for perception, the camera captures its field of view. However, seeing admits of three stages, each carrying different ethical and technical weight (Table 3).

Table 3
Three stages of seeing

Stage	Description	Status in this work
The eye as record	Recording only; nothing transmitted	Feasible in the preliminary system
The eye as perception	Frames transmitted to the model	Demonstrated here (author only)
The eye as attention	Joint attention on a shared object	The core task of the next stage

The third stage corresponds to what Tomasello (1995) argued is central to development. Realising it requires knowing where within the field of view the model directed its attention. The paper's main finding is that this quantity is unavailable in a cloud configuration.

The two roles conflict when the learner falls silent in thought. A camera that keeps its gaze fixed on the learner serves perception well but is oppressive as expression. We adopt the principle that expression takes precedence, and perception is a by-product.

The camera is mounted on a tabletop tripod whose legs straddle the speakerphone, so that capture and vision occupy a single footprint on the desk (Figure 1). The lens centre sits approximately 50 cm above the desk surface (Figure 2). On a standard desk of around 70 cm, this places the lens roughly 120 cm above the floor, against a measured eye height of 130 cm for the seated author (height 175 cm), some 10 cm below the line of sight. The height was chosen deliberately. Equipment placed where it must be looked up at readily acquires the connotation of surveillance; equipment set markedly lower reads as a subordinate instrument. A gaze that is near enough to level, and just fractionally lower, places the device where it is neither an agent that appraises the learner nor a tool in the learner's service.

That non-humanoid robots such as Keepon do not feel imposing owes something to their not looking down at people. This system likewise sits at a height from which it neither looks up at the learner nor down, and in doing so functions as a design element for suppressing the observer effect. The choice is also the physical expression of the position argued for in our conclusion: not an agent that judges, but an instrument that does not judge.

Figure 1 The preliminary system in place. The PTZ camera stands on a tabletop tripod that straddles the speakerphone, so that the camera and microphone share a single footprint on the desk

Figure 2 Dimensions of the assembly. The lens centre sits about 50 cm above the desk surface, roughly 120 cm above the floor on a standard desk.

Implementation

Table 4 shows the equipment configuration. The YVC-300 was connected directly via USB; the PTZ Pro 2 was connected through a USB hub with its dedicated AC adapter. A control comparison confirmed that this difference in connection does not affect the acoustic measurements reported below.

Table 4
Equipment configuration

Component	Model	Role
Capture and playback	Yamaha YVC-300	Unworn capture, response playback
Visual input	Logitech PTZ Pro 2	Still-frame capture
Computer	Notebook PC (Windows)	Control and logging

Capture conditions were held fixed throughout: 48,000 Hz sampling, 16-bit quantisation, one channel, WASAPI, microphone gain at 100, and a speaker distance of 60 cm. Where conversion was needed, it was performed in code; the stream settings themselves were never altered. The goal was to keep the observation conditions identical across sessions.

Spoken dialogue used the Google Gemini Live API. A half-cascade model served for connectivity checks; the main configuration used a native audio model. Two considerations motivated this choice. First, audio is processed without passing through text tokens, so the input stage does not lose paralinguistic information. Second, it presents, in its purest form, the argument developed below: there is no intermediate representation to read. One implementation constraint deserves note: no Live API model supports text responses, and requesting one terminates the connection abnormally. We therefore used a path that returns a transcription alongside the audio response.

To prevent the system's own response audio from re-entering the microphone and being taken as user speech, we implemented half-duplex control. Transmission to the model halts while a response is playing and resumes 300 ms after playback ends; recording itself never stops; playback intervals are logged in metadata and can be excluded afterwards. The dialogue system thus accepts a half-duplex constraint, while the research record remains continuous and without gaps.

Visual input was implemented as follows. The camera opens once when the dialogue loop starts and stays open until it ends, since opening and closing it each turn adds substantial latency. When the user's utterance is transmitted, one frame is captured, scaled to 640×480, encoded as JPEG, and attached to the same turn as the audio. The system saves captured frames and records the filename of the image sent in that turn in the metadata. If frame capture fails, the turn proceeds with audio alone.

No continuous video is transmitted. Restricting each turn to a single still image keeps bandwidth and processing modest while making it possible to identify unambiguously which frame corresponds to which response. That correspondence proved indispensable for adjudicating the confabulations discussed below.

For each session we retained the audio waveform, the transcription of responses, the transmitted images, and metadata including timestamps. We retained measurements that were subsequently rejected rather than deleting them: rejecting hypotheses is among the principal results of this work, and the primary evidence for it needs to survive.

Preliminary Observation

All measurements reported here were made on 17 August 2026 under the conditions given above. Table 5 summarises the baseline characteristics of the recording environment.

Table 5

Baseline characteristics of the recording environment

Item	Value	Notes
Noise floor	−63.4 dBFS (range −62 to −66)	Consistent across three independent measurements
Appropriate input level	−22.4 to −24.0 dBFS	No clipping; S/N 41 dB
Silence truncated by noise gating.	None observed	—
Convergence curve of adaptive DSP	Not observed	Total change −0.18 dB

That silence is not truncated is an important confirmation for this research. Time spent thinking in silence is itself an object of observation; if the device removes it, we cannot measure it. The absence of an observable DSP convergence curve is a negative result: under these conditions, we could not determine whether adaptive processing was active.

After capture, we recorded the difference between normal speech and deliberately quiet speech as 6.4 dB (n = 1). We infer that some compression is applied to the physical difference in vocal effort, but isolating the contribution of automatic gain control quantitatively would require simultaneous recording with a reference microphone. We do not assert compression here.

Steady-state suppression by the echo canceller measured −63.9 to −66.3 dBFS (n = 8), at or below the noise floor. Intermittent leakage was nonetheless observed. The leakage rate varied roughly sevenfold across source materials, from 7.7% to 51.8%, while remaining highly reproducible within a given source. Leakage is source-dependent; a leakage rate from a single source cannot be treated as a device performance figure. We could not identify the mechanism: playback level, onset steepness, playback duration, and hardware configuration each proved insufficient on its own.

Table 6

Response time and half-duplex behaviour

Condition	Measured	n
Half-duplex resumption	0.301 s	2
Response time (with speech)	2.10–4.66 s	—
Response time (silent input)	10.29–10.46 s	2

These response times cannot be used to compare conditions. The current implementation transmits audio in a single batch after the utterance ends, which differs from the incremental transmission a deployed system would use.

Dialogue with visual input was established. Frame capture took 7.3–16.6 ms (33.4 ms on the first call only), negligible against the response time; holding the camera open contributes here. Comparing the response against the transmitted frame, every element matched. Notably, the model read back a string printed on the subject's shirt, information specific to that frame, which could not have been produced from general knowledge or inference. This constitutes direct evidence that the visual input was in fact processed.

Six hypotheses were rejected against measurement in the course of this work (Table 7). The value lies in establishing, against the data, that no simple account suffices.

Table 7
Hypotheses rejected against measurement

Hypothesis	Grounds for rejection
The initial attenuation reflects DSP gain convergence	Level 28 dB above the actual noise floor; an order-of-magnitude irreconcilable difference. Traced to attenuation of the real sound
The echo canceller converges within 1.5 s	Leakage recurred mid-session across four sessions
Shorter responses leak more readily.	Correlation weak (r = −0.349) and non-monotonic
High far-end level produces leakage.	Two intervals at equal level, only one of which leaked
A steep onset following brief silence triggers leakage	No leakage even at an onset of +43.8 dB; leakage at positions with no onset at all
Cross-correlation can discriminate leakage.	Known echo (0.075–0.124) indistinguishable from noise floor (0.067–0.159); method withdrawn

Practical requirements that emerged during measurement are set out below. These are not merely procedural: they define what will be asked of the learner. Discard the first 15 seconds after the recording is started, to exclude keystroke noise. Do not leave the seat during recording; this is the principal source of transient noise. Exclude mobile phones from the room, or power them off. Compute the noise floor reference as a median, since transients pull means.

Limitations: What the Cloud Cannot Provide

This section is the heart of the paper. The preliminary system described above works; but as an instrument for the research proposed here it is inadequate in principle. We set out why, stage by stage.

The YVC-300 provides five kinds of signal processing: adaptive echo cancellation, noise reduction, automatic microphone tracking, automatic gain control, and reverberation suppression. For conferencing, these are useful. For research, they are a problem, because none of their parameters is exposed and none can be disabled. The compression of vocal effort noted above is, we infer, the work of automatic gain control, but since it cannot be switched off at the device, isolating it requires a separate reference microphone. The quantity we set out to observe has already been altered inside the device, before observation begins, and we have no means of knowing the nature or extent of that alteration.

The log probabilities a cloud API returns pertain to the model's own output. This research requires scoring the learner's utterance, which is a different quantity, and APIs that score arbitrary text are effectively unavailable. Hidden representations and attention weights are likewise undisclosed. Moreover, in a native audio model, where audio is processed without passing through text tokens, the intermediate representation one would want to read does not exist in the first place. The constraint is consistent across the chain (Table 8).

Table 8
Consistency of the constraint across the observation chain

Layer	Processing occurs	But
Capture (YVC-300)	Five DSP functions	No parameters exist; nothing can be turned off
Generation (cloud LLM)	Audio processed directly	No intermediate representation is returned

These constraints are not merely an abstract inconvenience. In our measurements, they surfaced as concrete harm. Classifying twelve turns of dialogue, three confabulations were observed. In the seven turns where speech was present at an appropriate level, −22 to −28 dBFS, none occurred; all three arose among the five turns of silent input, −59 to −64 dBFS. The observed confabulations fall into three types, distinguished by where the error originates (Table 9).

Table 9
Three types of confabulations

Type		Origin of the error	Detectability
A. Confabulation in recognition		ASR generates words from noise	Detectable by monitoring input level
B. Fabrication of a modality		Asserts an input type never transmitted	Detectable against the transmission log
C. False continuity	Context memory mistaken for input on the current turn		Difficult to detect; the content itself is correct

Type B is the most serious. The recogniser had correctly determined that no input was present. Even so, the generative side produced a fully articulated description, a search screen, corporate information, a share price and a market summary of a visual input for which no transmission path had been implemented at all. A correct determination existed inside the system, but it did not appear in the output. What an external observer can reach is the output alone.

More telling still is the contrast in Table 10. Under near-identical conditions, the system behaved in two different ways. The difficulty is not the non-determinism as such, but that nothing in the output allows us to work backwards and say which of the two occurred. Such an account would require attention weights, and the system does not return them. In the visual experiment, we could adjudicate the confabulation by comparing it with the transmitted image, because an original was available to compare against.

Table 10
A confabulating turn and its control

	Turn with confabulation	Control turn
Input level	−61.5 dBFS	−62.21 dBFS
Recognition	No input detected	No input detected
Visual input	None	None
Response	Detailed description of a non-existent image	Appropriate request for repetition

Brought to bear on the purpose of this research, the point can be put as follows. This research requires knowing what in the learner's utterance the model responded to. In a cloud configuration, however, the output does not let us distinguish a response grounded in the learner's speech from one grounded in the model's own priors. In the visual experiment, comparison against the transmitted image settled the question; a learner's inner state affords no original to compare against. For anything used as an instrument, the inability to trace the grounds of a response is disqualifying.

That image-grounded dialogue was established at all is a positive result. The model accurately read a string within its field of view, confirming that it genuinely processed the visual input. What cannot be obtained is where within that field of view the model directed its attention in arriving at the description. The eye as attention, joint attention on an object shared with the learner, depends on precisely this quantity.

A final constraint bears on the research infrastructure itself: models behind a cloud API may be updated without notice. For longitudinal observation, a measuring instrument whose characteristics change outside the experimenter's knowledge is fatal. Data accumulated over months cannot be guaranteed to come from the same instrument. Not being able to freeze the weights means not being able to guarantee the instrument remained the same. In a single sentence: what this research seeks is not better performance, but ownership of the instrument.

Discussion and Conclusion

This paper reports the stage preceding a methodology that would read the internal states of a dialogue model as an instrument for measuring human cognition. We built a preliminary voice-and-vision dialogue system and evaluated its properties as such an instrument.

We established both the spoken dialogue loop and image-grounded dialogue. The recording environment's noise floor measured −63.4 dBFS against appropriate speech at −22 to −28 dBFS, a separation of more than 35 dB. Silence, which the learner's thinking occupies, was confirmed not to be removed at the device. The visual response accurately read a string within the field of view.

Against this, obstructions to observation ran consistently through every stage of the chain. In the capture layer, the signal-processing parameters are neither exposed nor disableable. In the generative layer, neither log-likelihoods for the learner's speech, nor hidden representations, nor attention weights can be obtained.

The constraint surfaced as responses describing, in detail, visual inputs that had never been sent, even though the recogniser correctly registered their absence. A control turns under near-identical conditions asked appropriately for repetition, and nothing in the output tells us which of the two will occur. In the visual experiment, comparison against the transmitted image settled the matter. A learner's inner state affords no such original. For anything used as an instrument, the inability to trace the grounds of a response is disqualifying.

We conclude that what this research requires is not better performance but ownership of the instrument: a platform on which the weights can be fixed, the internal quantities obtained, and the treatment of silence determined by the experimenter. The necessity follows not from considerations of performance but from what the act of observation demands.

We note, finally, that the outward form of this preliminary system may call to mind a surveillant artificial intelligence familiar from cinema. Its role is the opposite. This system is not an agent that judges, but an instrument that does not. The distinction carries practical weight in negotiating deployment in an educational setting.

Future Work

The following were not resolved by the measurements reported here: the factor behind the roughly sevenfold variation in echo leakage rate across source materials; the compression ratio attributable to automatic gain control, measurable through simultaneous recording with a reference microphone; whether adaptive DSP is active at all under these conditions; whether barge-in is supported, which remains unmeasured; deployment latency under incremental transmission; how often Type C arises in multi-turn dialogue; and the drive noise, response speed, and control granularity of the PTZ mechanism.

Type C cannot be adjudicated from the output, because the content itself is correct. This is a general property of dialogue systems that carry context memory, and a platform change will not dissolve it. If internal states were obtainable, however, it may become possible to separate attention directed at the current turn's input from attention directed at context memory. This is one reason this research requires a local platform.

In this paper, the PTZ mechanism acquired visual input. Its expressive use, mapping the model's internal prediction error onto bodily movement, has not been implemented. In a cloud configuration, with no internal state available to map, embodiment can respond only to surface signals such as the presence or absence of a reply. Coupling internal state to movement, so that instrument and observed constitute each other in a closed loop, is the central task of the next stage.

The subject of this paper was the author alone. Verification with learners requires ethical review. A configuration that transmits video externally carries particular weight, since learners' faces would leave the institution; this too argues for a local platform.

Acknowledgement

This work was supported in part by JSPS KAKENHI Grant Number 26K05986.

References

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies, 59*(1–2), 119–155.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6*(3), 271–315.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). MIT Press.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language, 140*, 1–11.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 1–8).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.

Knowles, M. S. (1980). *The modern practice of adult education: From pedagogy to andragogy* (Rev. ed.). Cambridge Adult Education.

Kozima, H., Michalowski, M. P., & Nakagawa, C. (2009). Keepon: A playful robot for research, therapy, and entertainment. *International Journal of Social Robotics, 1*(1), 3–18.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarisation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919).

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science, 8*(3), 162–166.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319.

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Lawrence Erlbaum Associates.