Research Note 002 • 15 November 2026

The Third Level

Improving the derivation and meaning of the tokens a model ingests and predicts, for behavioural analysis and communication.

By **Joshua Bardsley** / Founder & CEO, FBED Labs

Two readers, one sentence

A seller says, “How are you handling that today?”

Give the sentence to two careful readers and ask what it is. They agree at once: a question. Ask what the seller was *doing* with it and the agreement frays. One reader hears diagnosis, a test of whether a problem exists. The other hears rapport, a polite way to keep the buyer talking. A third might call it the set-up for a pitch.

Every word arrived intact. What the sentence did is still in dispute.

We have seen this pattern repeatedly in our own reading work, including when the readers are frontier language models. Independent readers agree most readily on the *form* of a turn: a question, an acknowledgement, a statement. They agree least on its commercial *function*: whether the seller uncovered a problem, built credibility, or moved the deal forward. The closer a label gets to what a conversation is for, the less two readers share it.

I think that gap explains something many people building with AI have felt and few have named. Model capability has risen faster than almost anyone predicted, and commercial results have not risen with it. The models are not failing at language. They are being asked to work at a level of communication that language alone does not measure, using units nobody has yet shown two readers can agree on.

This note is about those units: how the tokens a behavioural model ingests and predicts should be derived, what makes them mean something, and why getting them right is the foundation the rest of commercial AI is missing. It ends with Shannon Nyquist, the model we are building to produce them, and with the environments and task sets we intend to release so others can test decisions against the same evidence.

1. Three problems of communication

In 1948 Claude Shannon published the mathematics of communication. A year later Warren Weaver wrote an introduction to it that divided the problem into three levels. [1,2]

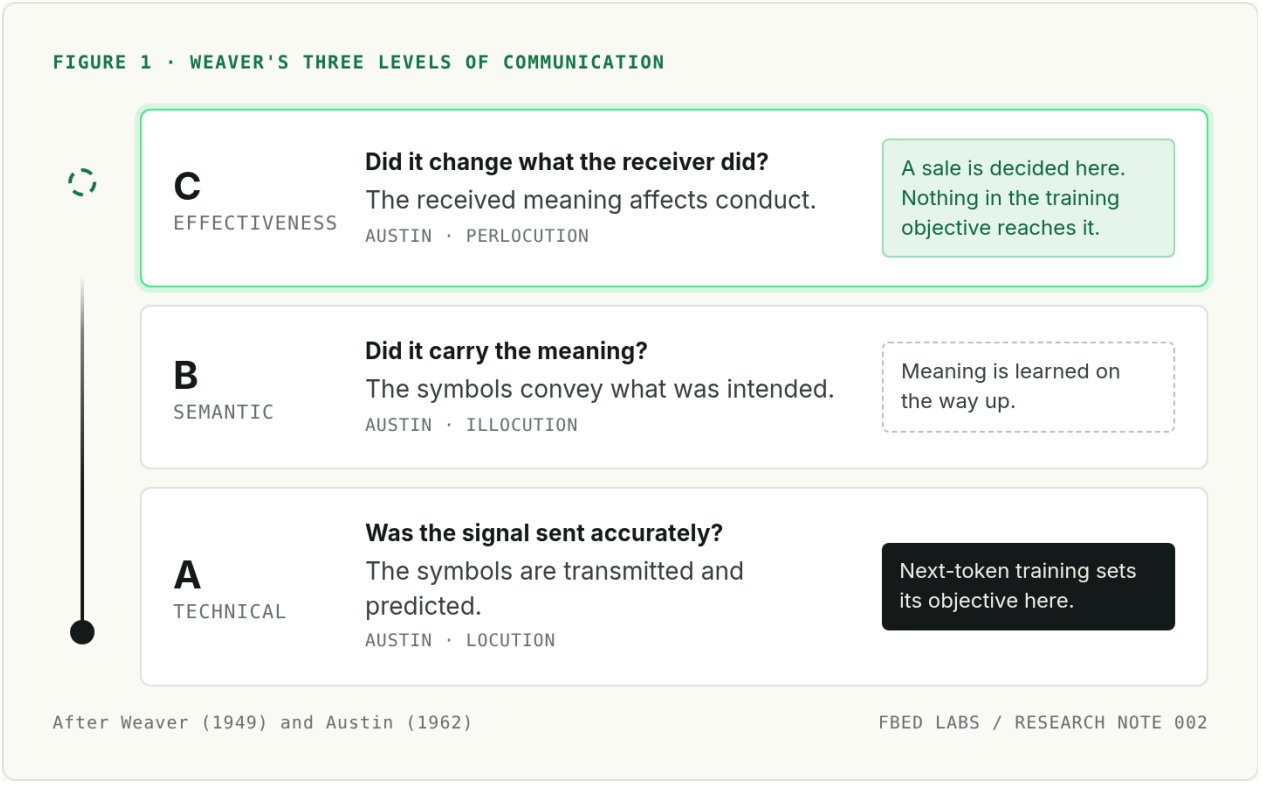
- **Level A, technical:** how accurately can the symbols be transmitted?
- **Level B, semantic:** how precisely do the symbols convey the intended meaning?
- **Level C, effectiveness:** how effectively does the received meaning affect conduct in the desired way?

Shannon's theory solved Level A, and deliberately so. “These semantic aspects of communication are irrelevant to the engineering problem,” he wrote. [1] Weaver's contribution was to insist that the other two

levels remain, and that a theory of the first might one day help with them.

Philosophy of language reached a similar ladder from the other side. J. L. Austin distinguished what is said, the **locutionary** act; what is done in saying it, the **illocutionary** act; and what is brought about by saying it, the **perlocutionary** effect.[3] A question is a locution. A diagnosis is an illocution. A buyer who reconsiders their problem is a perlocutionary effect.

A sale is a Level C problem. Nobody buys because a message was transmitted accurately, or even because it was understood. They buy, or don't, because of what the message changed in them and in the decision around them.



Next-token training sets its objective on the bottom rung. Meaning is learned on the way up. Nothing in the objective reaches the top.

2. What a language model learns to predict

A language model is trained to predict the next token in a sequence of text. Its objective, **cross-entropy** over tokens, is defined entirely at Level A: how well the model anticipates the symbols that come next.[4]

Meaning emerges from that objective, and remarkably well. A model that predicts text accurately enough must represent a great deal about what the text is about. Level B arrives as a by-product of Level A, one of the real surprises of the past decade.

Level C does not arrive. Nothing in next-token prediction says what the reader of a message did afterwards. Later stages of training add human preference: raters choose between answers and the model learns to produce the preferred one.[5] That moves the objective closer to a receiver, but the receiver is a rater judging a response, not a buyer changing a decision. Approval is not conduct.

This is why capability and commercial result can diverge without anyone being wrong. Benchmarks measure what models are trained to do, and they are improving fast. A commercial outcome is measured somewhere else, in what a counterpart did days or months later. When a quantity we can measure stands in for one we care about, and we optimise the stand-in, the two drift apart. That is **Goodhart's law**. The formal name for the underlying question, whether a measure captures the thing it claims to, is **construct validity**.[6,7]

Optimising for Level C is not new. Advertising has optimised clicks and conversions end to end for two decades, with no theory of dialogue at all. It works because the reward is dense, fast and attributable: millions of impressions, an answer within seconds, a click that belongs to one advert. A complex sale is the opposite. A handful of decisions, made by several people over months, with outcomes that are slow, sparse and shared between many causes. When reward is that scarce, what a model is allowed to observe matters far more.

The usual responses are more prompting, more context, a larger model. In our experience each helps at the margin, while the engineering needed for consistent behaviour on commercial communication climbs steeply. I think that is the signature of working at the wrong level: improving the channel when the problem sits at the receiver.

3. Tokens are a choice

Every model sees the world through an alphabet. For a language model it is a vocabulary of subword tokens, chosen by an algorithm such as **byte-pair encoding** to compress text efficiently.[8] It is a good alphabet for predicting text. It was never chosen to carry what a seller did or how a buyer responded.

A behavioural model needs a different alphabet: units of action and response. Linguistics has worked on this for decades under the name **dialogue acts**, labels such as question, acknowledgement or agreement attached to turns of conversation.[9] Those labels are useful, and they mostly describe form. The commercial functions that matter (uncovering a problem, building credibility, moving a deal forward) sit one level up and are harder to see.

The second choice is rate. Harry Nyquist showed in 1928, and Shannon proved in general in 1949, that a signal can be reconstructed from samples only if it is sampled at least twice as fast as it changes.[10,11] Sample more slowly and the signal is not merely blurred. It is misrepresented: a fast change appears as a slower, false one. Engineers call this **aliasing**.

Conversation is not the continuous, band-limited signal the theorem assumes, but by analogy behaviour aliases too. A buyer's trust can shift within a single exchange. A deal's stage may change only across weeks. A field recorded once per meeting cannot measure something that changes once per turn; at that rate it tells a smooth story about a jagged process. So every observation in a behavioural alphabet has to declare the rate at which it is sampled, and that rate has to be fast enough for what it claims to observe.

The tokens a model ingests are therefore a design decision with two parts: what each unit means, and how often it is read. Most commercial systems make both decisions implicitly. We think both should be made explicitly, and tested.

4. How a token earns its place

A label is not a measurement because someone wrote it down. Before a behavioural token can be trusted, it has to pass the tests that content analysis and psychometrics developed for exactly this problem.[12,7]

The first is **reliability**. If two readers, working independently from the same definitions, cannot agree that a turn contains a given act, the label does not yet refer to anything stable. Agreement has to be corrected for chance: two readers who both mark a rare label absent on almost every turn will agree almost all the time while sharing almost nothing. Statisticians call this the **prevalence paradox**, and it is why raw agreement flatters rare labels.[14] **Krippendorff's alpha** is one standard correction. Zero means agreement no better than chance; one means perfect agreement.[13]

Not every disagreement is error. Some turns really do two things at once, and a growing literature argues that disagreement between annotators can carry information of its own.[21,22] So readers should report a probability for each label rather than a single forced choice, and reliability should be measured over those probabilities. A turn that is ambiguous should be recorded as ambiguous, not settled by whichever reader happened to be more confident.

The other tests follow from the same idea.

- **Stability:** the same reader, reading again, reaches the same label.
- **Invariance:** a paraphrase that keeps the meaning keeps the label.
- **Sensitivity:** a near miss, a sentence that looks similar but does something different, changes it. Linguists call such pairs **minimal pairs**.
- **Calibration:** on constructed items whose answer is sealed in advance, the reader finds it.
- **Accuracy against gold:** finally, readers match an expert who knows the domain.

Reliability is not validity. Two readers can agree perfectly and both be wrong. But unreliability caps validity: the noise in a label limits how strongly it can predict anything, an effect psychometrics calls

attenuation.[12,20]

Our working rule is simple. A token that has not passed these tests is a hypothesis, and it is carried as one. It may be useful for exploration. It does not enter training as though it were an observation.

This is where the gap from the opening lives. Labels of form pass these tests readily. Labels of function, the ones a seller would recognise as moves, are where definitions blur, readers disagree and the work is hardest. They are also where the commercial value is.

One more discipline follows from Austin. He noted that an illocution depends on **uptake**: whether a question worked as a diagnosis depends partly on how it was received.[3] It is tempting, then, to label a turn by how it landed. Doing so lets the effect leak into the label of its cause, and a model trained on such labels learns that a move works partly because the move was defined by its working.

So each behavioural token comes in two halves. The **attempt** is what the seller tried to do, read from the turn and what came before it, blind to the reply. The **uptake** is how the buyer received it, read from the reply. Kept apart, they become exactly what a decision model needs: an action, and the observation that follows it.

5. Meaning is what a token predicts

Reliability tells us a token is observed consistently. It does not tell us the token matters.

For that we borrow a second idea from information theory. The **information bottleneck** asks for the most compressed description of an input that still keeps what the input says about a chosen target.[15] Applied to conversation, the input is the turn and its context, the target is what happens next, and a good behavioural alphabet is the compression that discards the most text while losing the least of the future. An alphabet should be shaped by what it has to carry, not by what is easy to name.

What a behavioural alphabet has to carry is the future. A token is meaningful, in the sense that matters here, to the extent that knowing it changes what we expect next: the buyer's next response, the next stage, the eventual decision. Physicists have a name for the part of the past that tells us about the future, **predictive information**.[16] Decision theory has a name for its worth, the **value of information**: the improvement in a decision that a piece of information makes possible.[17]

In practice the model's task is to predict uptake from the state and the attempted move: the next turn, and the turns after it. Predicting the future is the objective. Seeing it at the moment of decision would be cheating. Labelled uptake from past conversations is what such a model learns from, never what it is shown when it decides.

That gives a test harder than reliability. A behavioural alphabet should predict commercial transitions better than the raw text it was derived from, better than labels of form alone, and better than generic text

embeddings. If it cannot, it is a description, perhaps an elegant one, but not yet the physics of anything.

It also gives a definition of a good message that sits at Weaver's third level. A message is valuable to the degree that it reduces the receiver's uncertainty about something they need to decide. We call this **receiver-side value**. The words are the carrier. The change in the receiver is the value.

6. Shannon Nyquist: the base case

Shannon Nyquist is the model we are building to produce that alphabet. Its job is narrow and foundational: to sample a continuous conversation into discrete, typed observations at declared rates (per turn, per exchange, per meeting), so that the systems downstream reason over observations and not raw text.

It sits at the start of a family. Nyquist observes. Shannon Huffman evaluates and shapes a candidate outbound message against what those observations predict. Shannon Zero, described in *The Physics of a Sale*, plans over the commercial state they make possible.[18] Its problem has a formal name, a constrained Markov decision process: choose the seller's moves so that the commercial state moves toward a target while declared limits stay under their thresholds.[23] It steers the seller's choices. It does not control the buyer. Each depends on the one before it. A planner searching over unreliable observations is searching over noise, however good the search.

The base case for Nyquist rests on four claims.

1. **Every commercial AI system already consumes behavioural observations, whether it names them or not.** Scoring, coaching, forecasting, agent evaluation and outbound writing all decide what happened in a conversation before they decide what to do. Most make that decision implicitly, inside a prompt, differently each time.
2. **An explicit observation layer turns that hidden step into shared infrastructure.** If it is versioned and measured for reliability, two systems that read the same turn the same way can be compared. Two that do not cannot.
3. **Reliable function-level tokens make the commercial questions answerable.** Which move was made, how often, in what state, with what result. Without them, off-policy evaluation and the randomised comparisons described in our first note have nothing stable to count.[18]
4. **The observation layer compounds.** Models will be replaced. A well-defined, well-measured account of what happened in a conversation outlasts every model that reads it.

Conversation intelligence products already tag calls with topics, keywords and trackers. Nyquist differs in what it commits to for every observation: a measured and reported reliability, a declared sampling rate, a fixed version, and a definition sharp enough to be scored against what happened next.

The strongest objection is Rich Sutton's **bitter lesson**: hand-built representations tend to lose to general methods that learn from scale.[24] We take it seriously, and it is why the second claim in §8 compares our alphabet against generic embeddings. But scale wins where reward is dense and data abundant, and complex sales have neither. When outcomes are scarce, a model needs every piece of structure it can be given. An explicit alphabet is also what makes a policy auditable: even a system that learns its own representation has to be evaluated in units people can check.

Nyquist is released under the same discipline as the rest of our work. A version is minted only when its vocabulary, parameters, evaluation gates and runtime contract can be named and hashed, and its reliability is reported for every family of tokens alongside it. Until then it remains a research instrument.

7. Environments and task sets for decisions

Reinforcement learning moved faster once researchers shared a way to build **environments**: something that can be reset, that accepts an action, and that returns an observation and a reward. OpenAI Gym made that interface a common language.[19] Agent benchmarks have multiplied since. Most still grade at Weaver's first two levels: was the output correct, was it well formed, did a judge prefer it.

Commercial decisions have no environments of this kind. There is no shared place to ask whether an agent chose a better move for a given buyer in a given state, judged by what such buyers actually did.

We intend to release them. As Nyquist matures, we will publish environments and task sets built on its observations, designed to support decisions rather than text generation:

- **Environments** where the state is an evidence-backed commercial state, the actions are moves rather than strings, and the transitions are learned from observed behaviour, with their uncertainty stated.
- **Task sets**, each with a declared grader, a declared horizon and a **holdout** that nobody tunes against.
- **Reliability reports** for every observation an environment exposes, so a researcher knows which signals are firm and which are still hypotheses.

What we will not release matters as much. No transcript text, no client identities, and nothing that lets a counterparty be re-identified. The environments carry the structure of commercial behaviour, not the people in it.

A simulated buyer is still a prediction, as our first note argued. These environments are tools for comparing decisions under stated assumptions. The claim that a decision improves real outcomes is earned somewhere else: in a randomised comparison against a declared baseline, on real conversations, with the result kept whether it flatters us or not. We intend to run that comparison and publish it.

8. What would prove us wrong

A thesis is useful only if the evidence can contradict it. Ours makes three claims that can fail, in order.

1. **Function-level behavioural tokens can be made reliable.** If, after definitions are sharpened with minimal pairs and readers are checked against expert gold, independent readers still cannot agree on what a seller did, then commercial function may not be observable from conversation alone at the precision we need.
2. **Reliable tokens predict commercial transitions better than the alternatives.** If raw text or generic embeddings predict buyer responses as well as our alphabet does, the alphabet adds interpretability but not information, and the case for Nyquist narrows to that.
3. **Decisions made with those tokens improve outcomes.** If a policy using them does no better than the baseline in a randomised comparison, we will have built an accurate description of selling, and nothing more.

Each claim has to survive before the next is worth testing. The ordering is deliberate. It is far cheaper to discover that a label is unreliable than to discover, after a year of training, that a model learned to predict a label nobody could agree on.

There is also a question the method cannot settle on its own. A system that predicts uptake well could be pointed at persuasion for its own sake. Our first note set the line: the aim is not to control buyers, and a good system should sometimes help a seller see sooner that there is no sale. The limits a planner works under are where that commitment has to be written down, not left to good intentions.

Language models have given us machines that work fluently at the first two levels of communication. The third level, where a message changes what someone does, is where commerce lives, and its measurement has barely begun. We think it can be measured, starting with the units.

A token is a claim about what happened. Our job is to earn the right to make it.

Research note. This essay presents a research thesis and established scientific methods. It reports no quantitative FBED Labs measurements, and it does not claim that Shannon Nyquist is a deployed or independently validated model.

Glossary · plain words, formal names

In plain words	The formal term	First explained
Was the signal sent accurately?	**Level A, the technical problem**	§1
Did it carry the meaning?	**Level B, the semantic problem**	§1
Did it change what the receiver did?	**Level C, the effectiveness problem**	§1
What is said, what is done in saying it, what it brings about	**Locution, illocution, perlocution**	§1
How well a model predicts the next piece of text	**Next-token prediction; cross-entropy**	§2
Optimising a stand-in until it stops tracking the target	**Goodhart's law**	§2
Whether a measure captures what it claims to	**Construct validity**	§2
The units a model reads and predicts	**Tokens; tokenisation**	§3
Labels for what a turn of conversation does	**Dialogue acts**	§3
Sampling too slowly to see a change, so it looks like a different one	**Aliasing; the Nyquist rate**	§3
Whether independent readers give the same label	**Inter-rater reliability; Krippendorff's alpha**	§4
High agreement on a rare label for the wrong reason	**The prevalence paradox**	§4
Similar sentences that do different things	**Minimal pairs**	§4
Noise in a label limiting what it can predict	**Attenuation**	§4
What the seller tried, and how the buyer received it	**Attempt and uptake**	§4
The most compressed description that keeps what predicts the target	**Information bottleneck**	§5
The part of the past that tells us about the future	**Predictive information**	§5
How much better a decision gets with a fact	**Value of information**	§5
Planning toward a target while keeping declared limits	**Constrained Markov decision process**	§6
General methods that learn from scale beat hand-built structure	**The bitter lesson**	§6
A world that can be reset, takes an action and returns a result	**Environment**	§7

In plain words	The formal term	First explained
Data set aside that nobody tunes against	**Holdout**	§7

Research references

These sources establish the scientific methods discussed in the essay. They do not independently verify FBED Labs' data, experiments, implementation or commercial performance.

1. Shannon, C. E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal*, 27(3), 379–423 and 27(4), 623–656. DOI
2. Weaver, W. (1949). "Recent Contributions to the Mathematical Theory of Communication." In Shannon, C. E. & Weaver, W., *The Mathematical Theory of Communication*. University of Illinois Press.
3. Austin, J. L. (1962). *How to Do Things with Words*. Clarendon Press.
4. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. (2003). "A Neural Probabilistic Language Model." *Journal of Machine Learning Research*, 3, 1137–1155. JMLR
5. Ouyang, L. et al. (2022). "Training Language Models to Follow Instructions with Human Feedback." arXiv:2203.02155
6. Manheim, D. & Garrabrant, S. (2018). "Categorizing Variants of Goodhart's Law." arXiv:1803.04585
7. Cronbach, L. J. & Meehl, P. E. (1955). "Construct Validity in Psychological Tests." *Psychological Bulletin*, 52(4), 281–302. DOI
8. Sennrich, R., Haddow, B. & Birch, A. (2016). "Neural Machine Translation of Rare Words with Subword Units." *Proceedings of ACL 2016*. arXiv:1508.07909
9. Stolcke, A. et al. (2000). "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech." *Computational Linguistics*, 26(3), 339–373. ACL Anthology
10. Nyquist, H. (1928). "Certain Topics in Telegraph Transmission Theory." *Transactions of the AIEE*, 47(2), 617–644.
11. Shannon, C. E. (1949). "Communication in the Presence of Noise." *Proceedings of the IRE*, 37(1), 10–21. DOI
12. Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*, fourth edition. SAGE.
13. Hayes, A. F. & Krippendorff, K. (2007). "Answering the Call for a Standard Reliability Measure for Coding Data." *Communication Methods and Measures*, 1(1), 77–89. DOI
14. Feinstein, A. R. & Cicchetti, D. V. (1990). "High Agreement but Low Kappa: I. The Problems of Two Paradoxes." *Journal of Clinical Epidemiology*, 43(6), 543–549.
15. Tishby, N., Pereira, F. C. & Bialek, W. (1999). "The Information Bottleneck Method." *Proceedings of the 37th Allerton Conference on Communication, Control and Computing*. arXiv:physics/0004057
16. Bialek, W., Nemenman, I. & Tishby, N. (2001). "Predictability, Complexity, and Learning." *Neural Computation*, 13(11), 2409–2463. DOI
17. Howard, R. A. (1966). "Information Value Theory." *IEEE Transactions on Systems Science and Cybernetics*, 2(1), 22–26. DOI
18. Bardsley, J. (2026). *The Physics of a Sale: How Commercial Decisions Become Observable, Testable and Learnable*. FBED Labs Research Note 001. Read on fbed.io
19. Brockman, G. et al. (2016). "OpenAI Gym." arXiv:1606.01540
20. Spearman, C. (1904). "The Proof and Measurement of Association between Two Things." *American Journal of Psychology*, 15(1), 72–101. DOI
21. Plank, B. (2022). "The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation." *Proceedings of EMNLP 2022*.

22. Aroyo, L. & Welty, C. (2015). “Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation.” *AI Magazine*, 36(1), 15–24.

23. Altman, E. (1999). *Constrained Markov Decision Processes*. Chapman & Hall/CRC.

24. Sutton, R. S. (2019). “The Bitter Lesson.” Essay