Facilitating the Practical Implementation of Improved Explainability and Visual Representation for Confidence and Uncertainty in Speaker Models

Summary White Paper

Prepared May 1, 2025 by:
Helen Armstrong (PI), Matthew Peterson (co-PI), Rebecca Planchart, and Kweku Baidoo
Graphic and Experience Design, North Carolina State University

CONTENTS

1. Project Definition

Problem Statement. The Laboratory for Analytic Sciences (LAS) has established that there are significant challenges inherent to the calibration of trust within human-machine teams in the intelligence community (IC). The visualization of confidence and uncertainty, embedded within a user interface and user experience, should help language analysts appropriately calibrate trust via model transparency and interpretability. Such calibration could enable an analyst to more effectively evaluate model outputs when making a decision. Analysts should be able to “traverse different layers” within a user interface to access increasingly granular explanations of output (Knack et al., 2022, p. 5). If the user interface does not provide these explanations in a useful and usable format, analysts may distrust or overtrust model outputs (Lee & See, 2004, p. 73). To support the calibration of trust between analysts and speaker models, an effective visualization of confidence and uncertainty must be paired with a user interface and user experience that enable progressive disclosure of layered explanations as well as a dynamic system enabling analysts to adjust risk parameters in consideration of the larger mission context.

Research Question. How can interactive visualizations of confidence indicators enable language analysts to more accurately interpret and efficiently act on speaker model outputs?

Research Objectives. The project goal is to reveal potential innovations in visualization and interface design that will increase the likelihood of language analysts efficiently validating speaker model outputs, especially from a human-machine trust calibration perspective.

- **Objective 1:** Explore and evaluate *potential visualizations and UX patterns* for signifying confidence and uncertainty in speaker model outputs. [This involved collaboration with LAS experts and others in the intelligence community to ensure situational authenticity.]
- **Objective 2:** Create three different *visual prototypes* — in this case, mockups that provide explicit visual specifications for implementation — representing three possible solutions to this problem space. These visual prototypes should be structured so that usability testing might be efficiently conducted by the IC at the conclusion of the project.

2. Project Timeline

This investigation took place January 2025 through May 2025. Following are select design team activities conducted during that time.

In **January**, we:

1. Conducted interviews with language analysts about the use of the traffic search ("Tool 1") and media player ("Tool 2") interfaces.
2. Constructed a user journey map.
3. Identified nine key insights paired with user pain points.
4. Distributed additional surveys to language analysts.
5. Discussed current speaker model outputs with technical experts.
6. Developed the depth of engagement framework (see Section 3).
7. Discussed project parameters and file preparation with developers.
8. Demonstrated a Figma workflow to developers.

In **February**, we:

1. Refined the depth of engagement framework.
2. Surveyed 19 language analysts on their experience with confidence scores.
3. Created low-fidelity wireframes for Tool 2 options.
4. Developed 26 visualization strategies for indicating speaker model confidence.
5. Surveyed three language analysts on design progress.

In **March**, we:

1. Created two high-fidelity prototypes for indicating speaker model confidence in Tools 1 and 2.
2. Revised prototypes based on low-side analyst feedback.

3. Produced early prototype demonstration videos.
4. Surveyed 16 language analysts on prototype features and visualization strategies (A/B testing; see Section 6).
5. Consulted with developers on flagged design issues.
6. Created a feature table to maximize instructional variation among three final prototypes.
7. Revised three visualization strategies for indicating speaker model confidence (see Section 4)

In **April**, we:

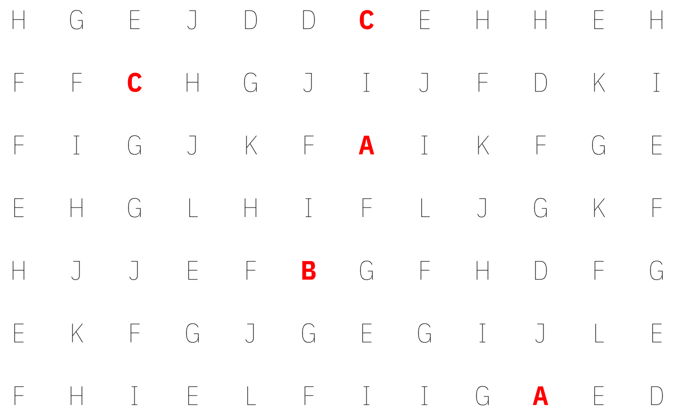
1. Created three interface prototype systems (across Tools 1 and 2).
2. Prepared markup for the development team (see Section 7).
3. Conducted an accessibility audit.
4. Produced final interface prototype scenario videos (see Section 5).

In **May**, we:

1. Presented the investigation and its outcomes to high side, low side, and LAS stakeholders.
2. Finalized the project deliverables, including this white paper.

3. Depth of Engagement Framework

3.1. Background. Spans of time as brief as milliseconds in the low hundreds (0.1–0.4s) are significant when they occur repeatedly in an analyst's workflow — when initially inspecting an interface or seeing search results appear. We found it useful to consider what kind of information can be immediately recognizable in a user interface. Though scene gist recognition in visual cognition (Oliva, 2005) typically considers naturalistic imagery (i.e., actual *scenes*), it suggests paying attention to discriminable features such as color for providing users with a roadmap for their unconscious gaze patterns as they determine what elements to engage with (or fixate on).



H	G	E	J	D	D	**C**	E	H	H	E	H
F	F	**C**	H	G	J	I	J	F	D	K	I
F	I	G	J	K	F	**A**	I	K	F	G	E
E	H	G	L	H	I	F	L	J	G	K	F
H	J	J	E	F	**B**	G	F	H	D	F	G
E	K	F	G	J	G	E	G	I	J	L	E
F	H	I	E	L	F	I	I	G	**A**	E	D

Figure 1. A field of arbitrary symbols with weight and color selectively applied to three kinds.

As a stark example, Figure 1 presents a field of the letters A–L. It is quite obviously much easier to count the total number of A's, B's, and C's than to count any other set of three letters. If the letters generally represent data points, and the highlighted letters are data points of interest — because they are uncertain, or because they signify an analyst's targets — then the clear signaling can make visual scanning far more efficient. Thus even peripherally visible elements can be top of mind.

The Figure 1 example is purely symbolic. Letters are arbitrary forms that must be learned, but when they *are* learned, their use becomes incredibly efficient. The color red in a field of black marks is also arbitrary, but its differentiation makes it distinctive. Peirce's pragmatic semiotics (Burks, 1949) provides two additional representational modes beyond the arbitrary *symbol*: the *icon,* which depicts or looks like the thing it represents; and the *index,* which is a visual form that has a correlational or causal relationship to the thing it (secondarily) represents. In Figure 2, the trap with cheese suggests a mouse, even though one is not pictured. Meaning for icons and indices is nowhere near as specific and controlled as it is for symbols (like written language), and in addition to being an index for *mouse,* the trap in Figure 1 is also an icon for... *trap.* Indices can be rooted in symbolism in addition to depiction — "squeak" is another index for *mouse.* This would be a largely useless classification exercise if understanding indices and symbols did not require viewer familiarity with dependent reference concepts. This matters for the selection of representations of confidence, since there is no thing *confidence* out in the world for us to show pictures of. Instead, we must use arbitrary symbols or find references that are useful without being misleading.

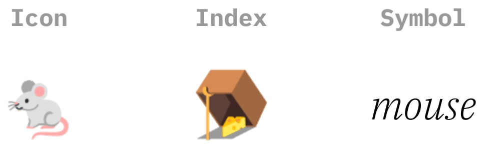


Figure 2. Icon, index, and symbol for the concept *mouse.*

Dual-coding theory (Sadoski & Paivio, 2001) differentiates between verbal and nonverbal representations, and the fact that humans have discrete mental resources for processing them means that their paired usage is generally additive, not a form of interference. We are thus motivated to dual-code important information when possible. Underlying these rudimentary considerations from theory is an interest in human information processing in the moments it takes analysts to find and assess estimations of confidence.

3.2. Premise. Informed by this background understanding, we determined a few basic assertions from which we derived a premise for our visualization strategy exploration.

A. Concepts can be represented through various schemes that differ semiotically — different schemes *mean* differently, and thus demand different things from users to achieve understanding.
B. Abstract concepts — such as *confidence* or *confidence score* — lack a direct corollary in the real world, making it impossible to determine "correct" representational accuracy. Thus evaluation must consider practicality, not accuracy.
C. Semiotic schemes can align with and thus emphasize inherent information features.
D. Semiotic schemes differ in how much information they can encode, and in the speed at which that information can be recognized by a user. (A depiction of a mouse is more direct than a mouse trap for representing *mouse.*)

The resultant premise is: *Semiotic schemes for confidence indicators should emphasize appropriate aspects of information uncertainty at appropriate moments in an analyst's workflow.* This emphasis on timing in workflow leads to our framework.

3.3. The framework. Our first insight when considering analyst workflow is relevant to all interface design: as a user visually inspects an information interface, different information becomes available earlier and later during inspection, which is mediated by representational format. In Figure 1, the letters A, B, and C are effectively *available* earlier than are the other letters despite being static. We can conceptualize information as existing at varying *depths of engagement,* based largely on its representational scheme. (In Figure 1, H is effectively *deeper* than B, since B is apparent more rapidly.)

Table 1. Single-screen depth of engagement scheme at four levels.

Shallower ↑	**Notice**	What is immediately available in the field of vision, even in the periphery, to inform eye movements.
	Read	What can be digested through more direct attendance to — or fixations on — visual elements.
	Probe	What is not immediately available but can be seen through noncommittal interaction (hover states).
↓ Deeper	**Inspect**	What is unavailable until there is a conscious commitment to accessing it (clicks, taps).

Table 2. Visual elements, visual characteristics, and interface patterns roughly mapped onto engagement depth levels.

Notice	**Read**	**Probe**	**Inspect**
Visual hierarchy Elevation Contrast (figure-ground) Animation Highlights White space	Textual content – explicit Numbers Labels Accordion (title – closed) Dropdown (title – closed)	Textual content – implicit Hover card Tooltip Status indicator Preview Quick access Scroll content – below the fold	Toggle Pop-up Lightbox Accordion (listing – open) Dropdown (listing – open) Filter (listing)

Table 1 outlines four engagement depths: *notice, read, probe,* and *inspect.* As levels get deeper, the user is making a stronger commitment to accessing information. The deeper levels in Table 1 require user interaction beyond eye movements. As shown in Table 2, different types of visual elements, visual characteristics, and interface patterns can be mapped onto the levels. This depth of engagement scheme permits the mapping of information at desirable instances in time. There may be some kind of information that is only important when users are inquiring in a special way. Such information may be distracting most of the time, and it can be pushed to a deeper engagement level to save it for special occasions.

Our second insight extends the depth of engagement scheme. We realized that the three tools (including Tool 3, data curation) are not themselves at the same level of engagement, but that they will tend to lead from one to the other in a typical workflow. The traffic search interface (Tool 1) will locate an audio cut for inspection in the media player (Tool 2). Table 3 presents this *continuous model* of the three interfaces, in which Tool 2's *notice* level is deeper than Tool 1's *inspect* level. An alternative *parallel model* is also presented, but the continuous model appears to better reflect use patterns in the analyst workflow.

Table 3. The depth of engagement framework. Engagement depths are mapped through Tools 1–3 ("Notice 1" refers to the *notice* level of Tool 1). The continuous model has 12 levels of depth in this scheme, and the parallel model four.

	Continuous Model	Parallel Model		
Shallower ↑	Notice 1			
	Read 1	Notice 1	Notice 2	Notice 3
	Probe 1			
	Inspect 1			
	Notice 2	Read 1	Read 2	Read 3
	Read 2			
	Probe 2			
	Inspect 2	Probe 1	Probe 2	Probe 3
	Notice 3			
	Read 3			
	Probe 3	Inspect 1	Inspect 2	Inspect 3
↓ Deeper	Inspect 3			

Tool 3 is beyond the scope of the project, and thus its levels can be collapsed to represent anything that belongs therein — and thus is not under consideration here. In this way excluding a potential information source in Tool 1 and Tool 2 prototypes does not necessarily imply a recommendation that the source should be inaccessible to analysts, only that it might be more appropriate at a level as deep as Tool 3.

The depth of engagement framework suggests some questions relevant to the project, including:

- What depth of engagement is appropriate for fulfilling the user's "right of explanation" for a given facet of intelligence analysis?
- What gradation of confidence indication is appropriate at a given depth of engagement? (0–100%: 101 levels; • ● ⬤: three levels; 🟨🟧🟪🟪🟦: five levels)
- At what depth of engagement is it helpful to represent model health, and at what depth is it distracting?

- At what depth of engagement is it helpful to present alternative speakers beyond the most likely speaker alone?
- At what confidence score threshold does it become helpful to present alternative speakers?

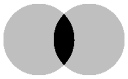
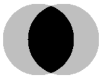
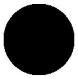
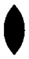
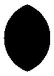
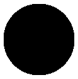
In this way the framework can guide or frame design decisions.

4. Confidence Visualization Strategies

4.1. Determination of schemes. Informed by our semiotic considerations and the depth of engagement framework, we developed 26 schemes for visually indicating confidence. In each case the underlying confidence scores of 0–100% (101 levels) would remain unchanged, with bands of percentage scores mapping onto (or collapsing into) visualizations, thereby reducing the total number of apparent levels of gradation. In no instance did we consider retaining the full 101 levels normally presented in the underlying confidence scores, because this granularity does not appear to align well with analyst decision making.

Through multiple rounds of feedback with various stakeholders as well as our own analysis, we reduced the number of visualization strategies under consideration to 10 and then eventually down to the three we have prepared for user testing. Table 4 provides an instructive example of a visualization strategy that we ultimately rejected.

Table 4. Confidence indication with a Venn diagram "matching" analogy.

	Confidence Level		
Version	**Low**	**Medium**	**High**
Deeper version shown for explanation			
Shallower version for compact usage			
Descriptor	Minimal Match	Moderate Match	Significant Match

The following points describe our understanding of this visualization strategy:

A. The deeper-shallower distinction permits lightweight representations en masse in a user interface (the monotone shallower versions) with visually-supplemented explanations for the scheme to pair with system descriptions or to use in analyst orientation (the two-tone deeper versions). This is a unique affordance of this particular visualization strategy — other strategies have their own dissimilar affordances. This is the nature of the slippery meaning of imagery, in contrast with the precision of symbolic natural language, that schemes can have added benefits and limitations.

B. The Venn diagram analogy is, in our estimation, the most conceptually accurate visualization strategy for the situation. One circle represents the profile (or vectors) of recorded speech, while the other circle represents the profile of a given speaker model. The greater the overlap between these two patterns, the higher the confidence in speaker attribution.

C. As demonstrated in [B], an explanation is necessary for the scheme to make sense. It also requires a baseline understanding of speaker model technology, further complicating matters. Thus, we anticipated this not being a successful visualization strategy.

D. Unsurprisingly as suggested in [C], surveyed analysts did not favor the scheme.

Arguably only [D] matters. A good visualization strategy in this context must be fairly immediate. Analysts have been using confidence scores for some time, and a new implemented scheme cannot require too much explanation given such continuity, as there is no controlled situation by which to predetermine users' conceptualizations of confidence indicators — their conceptualizations are established.

After soliciting stakeholder and user feedback, we settled on three visualization strategies: (A) Square Digits, (B) Arc Gauge, and (C) Bar Fill. Arc Gauge and Bar Fill utilize the same bracketing from source confidence scores. Our understanding of the existing confidence scores is that 100% is a special indicator for selected speech that a speaker model has been analyst-verified as correct. We incorporate this understanding into our score-bracketing patterns.

4.2. Visualization strategy A: Square Digits. The Square Digits scheme uses symbolic representations of single digits to represent levels of confidence (Table 5). It thus does not deviate significantly from the existing percentage scores semiotically, but it has greatly reduced levels of granularity.

Table 5. Square Digits visualization strategy for speaker model confidence. Alternate indicators are dependent upon a user setting; two settings are shown: 7+ and 9+. While source files are in vector format and are thus scalable, 18×18 pixel versions are shown here to demonstrate visual clarity.

Range	Indicator		Descriptor
100%	☑	☑	Analyst verified
90–99%	9	9	Maximum reasonable confidence
80–89%	8	8	Very high confidence
70–79%	7	7	High confidence
60–69%	6	6	Fairly high confidence
50–59%	5	5	Moderate confidence
40–49%	4	4	Somewhat low confidence
30–39%	3	3	Low confidence
20–29%	2	2	Very low confidence
–	☒	☒	Analyst refuted
–	👤?	👤?	Unidentified speaker
–	!	!	AI generated

Square Digits is dual-coded and unique in accommodating user settings for *notice*-level signaling of confidence level. As in Figure 1, the solid blue boxes are rapidly differentiated from the outlined boxes, enabling an analyst to determine the relative confidence level that stands out for them, while retaining the same underlying single-digit values within teams and the agency.

Square Digits is also unique in the degree to which it follows from the existing use of confidence scores, possibly making it easier for continuing analysts to adapt to. (Newly hired analysts would not have this benefit, and for them all schemes would be equally unfamiliar.) For this reason, we recommend that the indicated range brackets not be changed. They permit analysts familiar with the existing confidence scores to seamlessly convert their experiential understanding of those scores to the new indicators, because the digits are shorthand for the 10s digit place in existing scores.

The number of confidence levels indicated with Square Digits is a factor of base-10 numerics, not an indication that we believe this to be the “right” number of levels. The benefit of going up to “9” in a single digit scheme is that it implicitly suggests a ceiling level to users without the need for explanation.

Finally, Square Digits is unique in foregoing visual metaphor and analogy in utilizing numeric representations. This may make the scheme less meaningful to analysts initially, and/or it may make its use more precise and rapid long-term — this is the nature of learned symbols.

In all three selected visualization strategies, a qualitative indicator signifies analyst verification instead of co-opting the otherwise continuous “100%” as is currently done. Also in all versions, the same “!” and “X” icons are available for implementing warnings that given speech may have been AI generated, or that an analyst rejected an identified speaker, respectively. These are documented in Table 5. The verification check mark is an index, suggesting that somebody did the checking — a human, an analyst.

We do not include confidence indicators in the range of underlying 0–20% confidence scores because we believe the current system does not display speaker model results in this range. If this does however occur, then for Square Digits, “1” is in the range of 10–19%, and “0” is 0–9%.

4.3. Visualization strategy B: Arc Gauge. The Arc Gauge scheme uses a tachometer (or similar) metaphor to suggest lesser or greater, or less or more powerful, performance of a speaker model (Table 6).

Table 6. Arc Gauge visualization strategy for speaker model confidence. 18×18 pixel indicator versions are shown here to demonstrate visual clarity.

Range	Indicator	Descriptor
100%	●	Analyst verified
80–99%	◕	High confidence
45–80%	◑	Moderate confidence
20–45%	◔	Low confidence
—	○	Unidentified speaker

The tachometer metaphor is not highly articulated in that it only gives a general sense of relative value, without the metaphorical source and target domains directly aligning in analogous function. The scheme in

Table 4 demonstrated that doing this is not necessarily a path to success — sometimes users understand simpler schemes even without much conceptual depth. We do think that the *performance* concept is a sensible alignment, and it was not surprising that surveyed analysts responded positively to Arc Gauge. The descriptors for this scheme reflect our belief that the metaphor is not highly articulated — we recommend the plain language of "high confidence" to "low confidence."

Arc Gauge is dual-coded, with color supplementing the metaphorical imagery. The main color sequence utilizes the green, yellow, red traffic signal pattern as a parallel to high- to low-confidence.

The range bracketing of Arc Gauge (as well as Bar Fill) is reduced to three levels, in order to reduce analyst cognitive load and align granularity with consequent actions. The range brackets are derivative of ICD 203 Analytic Standards in utilizing breakpoints at 45% and 80%, which are shared with the likelihood breakpoints. However, based on stakeholder recommendations, we in turn do not recommend utilizing the ICD 203 terminology (e.g., "unlikely" or "improbable"), for fear of confusing the *speaker model confidence* construct with other constructs that analysts may routinely encounter.

The strength of the Arc Gauge scheme is that it has the benefits of an easily understood visual metaphor.

4.4. Visualization strategy C: Bar Fill. The Bar Fill scheme uses a rudimentary volume analogy to suggest an amount of certainty (Table 7). Its orientation may also call to mind a battery of varying charge, which we consider to be a complementary meaning that does not reduce the scheme's efficacy.

Table 7. Bar Fill visualization strategy for speaker model confidence. 18×18 pixel indicator versions are shown here to demonstrate visual clarity.

Range	Indicator	Descriptor
100%	☑	Analyst verified
80–99%	🟩	Excellent
45–80%	🟨	Good
20–45%	🟥	Fair
—	□	Unidentified speaker

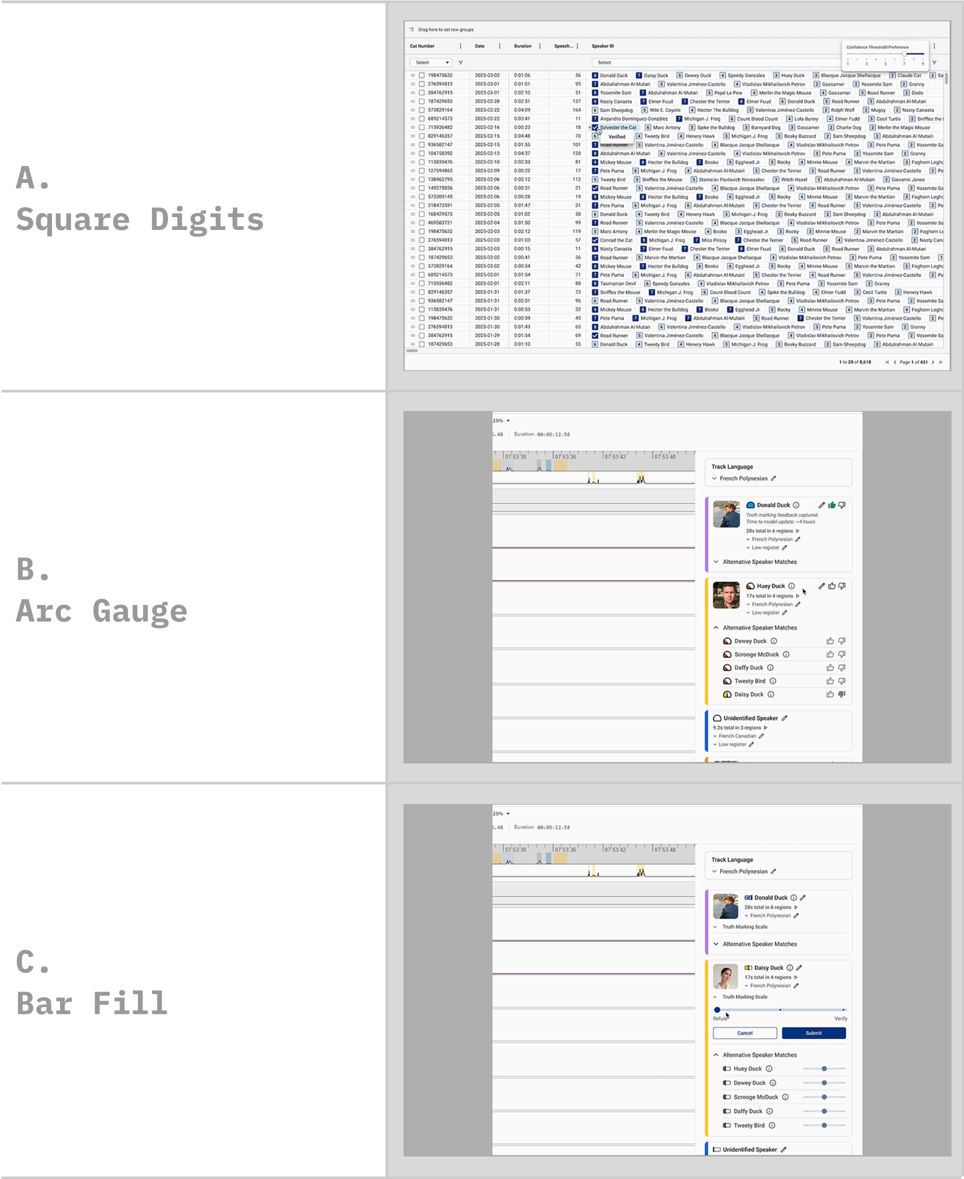
Like Arc Gauge, Bar Fill is both dual-coded with the green, yellow, red color scheme and aligned to ICD 203 breakpoints. It is unique among the visualization strategies in indicating an unidentified speaker as a continuation of certainty, with an empty bar effectively standing for "no confidence."

The strength of the Bar Fill scheme is its middle-ground position. It is an analogy that is more visually meaningful than the symbolic digits of Square Digits, and that does not have any potential negative metaphorical entailments given its semiotic simplicity. (This does not mean that we think the Arc Gauge metaphor is problematic. We are making note of *possible* interpretations of *potential* user test results.)

4.5. Visualization strategy summary. The three selected visualization strategies differ from one another semiotically, and Square Digits has implications for user interface design that the others do not share. The confidence score range bracketing listed for Arc Gauge (Table 6) and Bar Fill (Table 7) is an initial recommendation only — ideally the bracketing would be determined with data from practicing analysts, reflecting what is most useful to them. In contrast, the range bracketing listed for Square Digits (Table 5) is logically important, being related to the familiar percentage confidence scores currently in use.

5. Interface Prototypes

5.1. Scenario Videos. We produced separate scenario videos for each of three potential interfaces with distinct visualization strategies (Figure 3). They depict an interaction sequence beginning in Tool 1 and continuing into Tool 2.



Cut Number	Date	Duration	Speech...	Speaker ID
☐ 198475632	2025-03-02	0:01:06	56	8 Donald Duck 7 Daisy Duck 5 Dewey Duck 4 Speedy Gonzales 3 Huey Duck 3 Blacque Jacque Shellacque 2 Claude Cat 2 Sa
☐ 276594813	2025-03-01	0:01:01	95	7 Abdulrahman Al-Mutari 5 Valentina Jiménez-Castello 4 Vladislav Mikhailovich Petrov 2 Gossamer 2 Yosemite Sam 2 Granny
☐ 384762915	2025-03-01	0:02:10	51	8 Yosemite Sam 7 Abdulrahman Al-Mutari 5 Pepé Le Pew 4 Merlin the Magic Mouse 4 Gossamer 3 Road Runner 2 Dodo
☐ 187429653	2025-02-28	0:02:51	137	9 Nasty Canasta 7 Elmer Fuud 7 Chester the Terrier 8 Elmer Fuud 6 Donald Duck 5 Road Runner 3 Abdulrahman Al-Mutari
☐ 573829164	2025-02-22	0:04:09	164	6 Sam Sheepdog 5 Wile E. Coyote 4 Hector The Bulldog 3 Valentina Jiménez-Castello 2 Ralph Wolf 2 Mugsy 2 Nasty Canasta
☐ 689214573	2025-02-22	0:03:41	11	7 Alejandro Domínguez-González 7 Michigan J. Frog 6 Count Blood Count 4 Lola Bunny 4 Elmer Fudd 3 Cecil Turtle 2 Sniffles the
☐ 715936482	2025-02-16	0:00:23	18	✓ Sylvester the Cat 6 Marc Antony 5 Spike the Bulldog 3 Barnyard Dog 3 Gossamer 2 Charlie Dog 2 Merlin the Magic Mouse
☐ 829146357	2025-02-15	0:04:48	70	6 Verified 4 Tweety Bird 4 Henery Hawk 3 Michigan J. Frog 3 Beaky Buzzard 2 Sam Sheepdog 2 Abdulrahman Al-Mutairi
☐ 936582147	2025-02-15	0:01:55	101	7 Road Runner 5 Valentina Jiménez-Castello 4 Blacque Jacque Shellacque 4 Vladislav Mikhailovich Petrov 2 Pete Puma 2 Yosemite Sa
☐ 104758392	2025-02-13	0:04:37	128	8 Abdulrahman Al-Mutari 6 Valentina Jiménez-Castello 4 Vladislav Mikhailovich Petrov 3 Pete Puma 2 Yosemite Sam 2 Granny
☐ 115839476	2025-02-10	0:02:33	81	9 Mickey Mouse 8 Hector the Bulldog 7 Bosko 5 Egghead Jr. 5 Rocky 4 Minnie Mouse 4 Marvin the Martian 3 Foghorn Leghc
☐ 127594863	2025-02-09	0:00:22	17	7 Pete Puma 6 Michigan J. Frog 4 Abdulrahman Al-Mutairi 3 Chester the Terrier 3 Road Runner 2 Valentina Jiménez-Castello 2 Pe
☐ 138462795	2025-02-06	0:02:12	112	5 Tweety Bird 5 Sniffles the Mouse 5 Stanislav Pavlovich Novoselov 3 Witch Hazel 3 Abdulrahman Al-Mutairi 2 Giovanni Jones
☐ 149275836	2025-02-06	0:00:31	21	✓ Road Runner 5 Valentina Jiménez-Castello 4 Blacque Jacque Shellacque 4 Vladislav Mikhailovich Petrov 2 Pete Puma 2 Yosemite Sa
☐ 573209145	2025-02-06	0:00:28	19	9 Mickey Mouse 8 Hector the Bulldog 7 Bosko 6 Egghead Jr. 5 Rocky 4 Minnie Mouse 3 Marvin the Martian 3 Foghorn Leghc
☐ 218473591	2025-02-05	0:01:47	31	7 Pete Puma 6 Michigan J. Frog 6 Abdulrahman Al-Mutairi 5 Chester the Terrier 5 Road Runner 3 Valentina Jiménez-Castello 2 Pe
☐ 168429573	2025-02-05	0:01:02	58	6 Donald Duck 4 Tweety Bird 4 Henery Hawk 3 Michigan J. Frog 3 Beaky Buzzard 2 Sam Sheepdog 2 Abdulrahman Al-Mutairi
☐ 469583721	2025-02-04	0:01:50	99	7 Road Runner 5 Valentina Jiménez-Castello 4 Blacque Jacque Shellacque 4 Vladislav Mikhailovich Petrov 2 Pete Puma 2 Yosemite Sa
☐ 198475632	2025-02-03	0:02:12	119	5 Marc Antony 4 Merlin the Magic Mouse 4 Bosko 3 Egghead Jr. 3 Rocky 3 Minnie Mouse 2 Marvin the Martian 2 Foghorn Le
☐ 276594813	2025-02-03	0:01:03	57	✓ Conrad the Cat 8 Michigan J. Frog 7 Miss Prissy 7 Chester the Terrier 5 Road Runner 4 Valentina Jiménez-Castello 2 Nasty Cana
☐ 384762915	2025-02-03	0:00:15	11	9 Nasty Canasta 7 Elmer Fuud 7 Chester the Terrier 8 Elmer Fuud 6 Donald Duck 5 Road Runner 3 Abdulrahman Al-Mutari
☐ 187429653	2025-02-02	0:00:41	36	7 Road Runner 5 Marvin the Martian 4 Blacque Jacque Shellacque 4 Vladislav Mikhailovich Petrov 2 Pete Puma 2 Yosemite Sam 1
☐ 573829164	2025-02-02	0:00:54	42	8 Mickey Mouse 7 Hector the Bulldog 6 Bosko 6 Egghead Jr. 5 Rocky 4 Minnie Mouse 3 Marvin the Martian 2 Foghorn Leghc
☐ 689214573	2025-02-01	0:01:54	71	7 Pete Puma 6 Michigan J. Frog 6 Abdulrahman Al-Mutairi 5 Chester the Terrier 4 Road Runner 5 Valentina Jiménez-Castello 2 Pe
☐ 715936482	2025-02-01	0:02:11	88	8 Tasmanian Devil 6 Speedy Gonzales 4 Vladislav Mikhailovich Petrov 3 Pete Puma 2 Yosemite Sam 2 Granny
☐ 829146357	2025-01-31	0:01:37	73	7 Sniffles the Mouse 7 Michigan J. Frog 6 Count Blood Count 4 Spike the Bulldog 4 Elmer Fudd 3 Cecil Turtle 2 Henery Hawk
☐ 936582147	2025-01-31	0:02:01	96	6 Road Runner 5 Valentina Jiménez-Castello 4 Blacque Jacque Shellacque 4 Vladislav Mikhailovich Petrov 2 Pete Puma 2 Yosemite Sa
☐ 115839476	2025-01-31	0:00:53	32	9 Mickey Mouse 8 Hector the Bulldog 7 Bosko 7 Egghead Jr. 5 Rocky 4 Minnie Mouse 4 Marvin the Martian 4 Foghorn Leghc
☐ 198475632	2025-01-30	0:00:59	45	7 Pete Puma 7 Michigan J. Frog 7 Abdulrahman Al-Mutairi 5 Road Runner 7 Chester the Terrier 5 Valentina Jiménez-Castello 2 Pe
☐ 276594813	2025-01-30	0:01:43	63	8 Abdulrahman Al-Mutari 6 Valentina Jiménez-Castello 4 Vladislav Mikhailovich Petrov 3 Pete Puma 2 Yosemite Sam 2 Granny
☐ 384762915	2025-01-29	0:01:54	69	✓ Road Runner 5 Valentina Jiménez-Castello 4 Blacque Jacque Shellacque 4 Vladislav Mikhailovich Petrov 2 Pete Puma 2 Yosemite Sa
☐ 187429653	2025-01-28	0:01:10	55	6 Donald Duck 4 Tweety Bird 4 Henery Hawk 3 Michigan J. Frog 3 Beaky Buzzard 2 Sam Sheepdog 2 Abdulrahman Al-Mutairi

1 to 20 of 8,618 Page 1 of 431

Track Language
French Polynesian

Donald Duck
Truth marking feedback captured.
Time to model update: ~4 hours
28s total in 6 regions
French Polynesian
Low register
Alternative Speaker Matches

Huey Duck
17s total in 4 regions
French Polynesian
Low register
Alternative Speaker Matches
Dewey Duck
Scrooge McDuck
Daffy Duck
Tweety Bird
Daisy Duck

Unidentified Speaker
9.2s total in 3 regions
French Canadian
Low register

Track Language
French Polynesian

Donald Duck
28s total in 6 regions
French Polynesian
Truth Marking Scale
Alternative Speaker Matches

Daisy Duck
17s total in 4 regions
French Polynesian
Truth Marking Scale
Refute Verify
Cancel Submit
Alternative Speaker Matches
Huey Duck
Dewey Duck
Scrooge McDuck
Daffy Duck
Tweety Bird

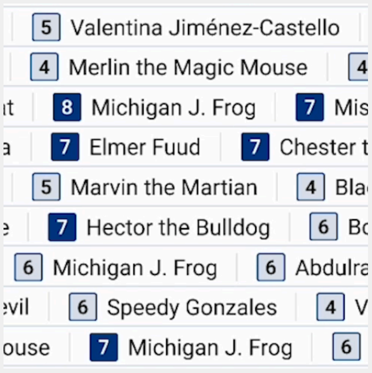
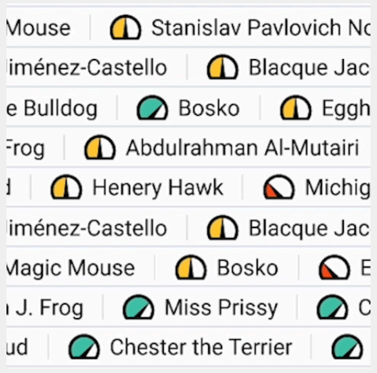
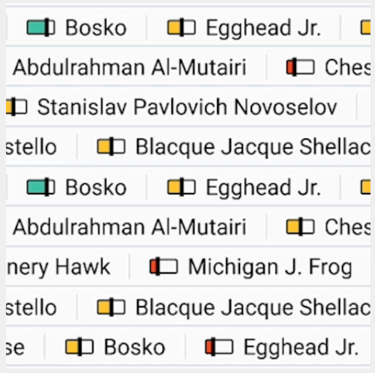
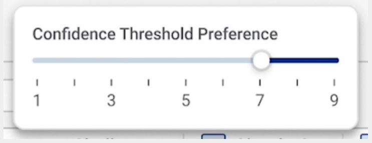
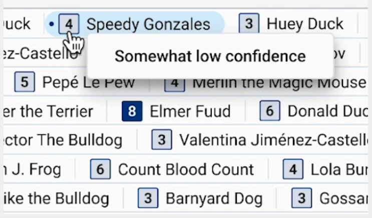
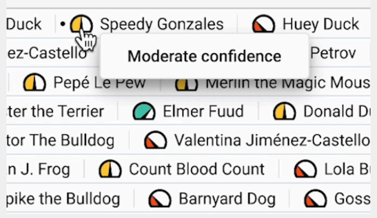
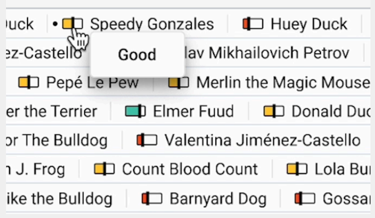
Unidentified Speaker

Figure 3. Interface prototype scenario videos.

5.2. Prototype Features. The scenario videos provide greater detail on potential implementation than we can outline here. Though we document many of the features and distinctions below, designers and developers who are acting on our recommendations should use the scenario videos to compare features and stylistic decisions across and among the three versions.

Tool 1 differentiation is largely restricted to the visualization strategies themselves, with the notable exception of the *confidence threshold preference* in the Square Digits version (Table 8).

Table 8. Tool 1 features in the three interface prototypes, with screenshots provided for comparison. ⋆ Insights are described in Section 6.

A. Square Digits Scheme	B. Arc Gauge Scheme	C. Bar Fill Scheme
 5 Valentina Jiménez-Castello 4 Merlin the Magic Mouse 4 it 8 Michigan J. Frog 7 Mis a 7 Elmer Fuud 7 Chester t 5 Marvin the Martian 4 Bla e 7 Hector the Bulldog 6 Bc 6 Michigan J. Frog 6 Abdulra evil 6 Speedy Gonzales 4 V ouse 7 Michigan J. Frog 6	 Mouse Stanislav Pavlovich Nc iménez-Castello Blacque Jac e Bulldog Bosko Eggh Frog Abdulrahman Al-Mutairi Henery Hawk Michig iménez-Castello Blacque Jac Magic Mouse Bosko E J. Frog Miss Prissy C ud Chester the Terrier	 Bosko Egghead Jr. Abdulrahman Al-Mutairi Ches Stanislav Pavlovich Novoselov stello Blacque Jacque Shellac Bosko Egghead Jr. Abdulrahman Al-Mutairi Ches nery Hawk Michigan J. Frog stello Blacque Jacque Shellac se Bosko Egghead Jr.
Square Digits visualization used	Arc Gauge visualization used	Bar Fill visualization used
 Confidence Threshold Preference 1 3 5 7 9	No visual confidence threshold setting available (Insights H and P*)	
Visual confidence threshold setting available (Insights H and P*)		
 uck • 4 Speedy Gonzales 3 Huey Duck ez-Castello Somewhat low confidence ov 5 Pepé Le Pew 4 Merlin the Magic Mouse er the Terrier 8 Elmer Fuud 6 Donald Duc ctor The Bulldog 3 Valentina Jiménez-Castell J. Frog 6 Count Blood Count 4 Lola Bur ike the Bulldog 3 Barnyard Dog 3 Gossar	 Duck • Speedy Gonzales Huey Duck ez-Castello Moderate confidence Petrov Pepé Le Pew Merlin the Magic Mous ter the Terrier Elmer Fuud Donald D tor The Bulldog Valentina Jiménez-Castello n J. Frog Count Blood Count Lola B pike the Bulldog Barnyard Dog Goss	 uck • Speedy Gonzales Huey Duck ez-Castello Good lav Mikhailovich Petrov Pepé Le Pew Merlin the Magic Mouse er the Terrier Elmer Fuud Donald Duc or The Bulldog Valentina Jiménez-Castello J. Frog Count Blood Count Lola Bur ike the Bulldog Barnyard Dog Gossar
Option to reveal confidence descriptor language upon click		

Differentiation is more significant in Tool 2, with some feature variation tracked in Table 9. Differentiated features may be used for A/B testing. In most cases, we have engineered variation among the interface prototypes to provide alternatives, not as a declaration that a given distinction is only appropriate for the visualization strategy that coincides with it — mixing and matching is encouraged, based on additional insights gained from any user testing, or upon implementation limitations.

Table 9. Tool 2 features in the three interface prototypes. ⋆ Insights are described in Section 6.

<table>
<tr><th>A. Square Digits Scheme</th><th>B. Arc Gauge Scheme</th><th>C. Bar Fill Scheme</th></tr>
<tr><td>Square Digits visualization used</td><td>Arc Gauge visualization used</td><td>Bar Fill visualization used</td></tr>
<tr><td>Visual confidence threshold setting carries over from Tool 1 (Insights H and P*)</td><td colspan="2">No visual confidence threshold setting available (Insights H and P*)</td></tr>
<tr><td colspan="3">Certainty description included via click on confidence indicator</td></tr>
<tr><td colspan="3">Truth marking history (total up-down arrows) accessible on hover via “i” symbol (Insight D*)</td></tr>
<tr><td colspan="2">Last person to truth mark indicated (Insight T*)</td><td>No indication of last person to truth mark (Insight T*)</td></tr>
<tr><td colspan="2">Truth marking input: thumbs up/down (Insight C*). Click a second time on thumbs up/down to deselect.</td><td>Truth marking input: three-point scale (Insight C*). Click on the central point to deselect.</td></tr>
<tr><td colspan="3">Time to model update (approx. 4 hrs) revealed upon <i>verify</i> or <i>refute</i> action (Insight E*)</td></tr>
<tr><td colspan="3">Option to view five alternative speaker matches (Insight G*)</td></tr>
<tr><td colspan="3">Option to edit the name of the top speaker choice, automatically verifying that choice. The previous name moves one spot down the alternative speaker list. The same action occurs when a speaker is verified without editing (Insight G*).</td></tr>
<tr><td colspan="3">Option to edit designated language indicator included on individual and whole cut level</td></tr>
<tr><td colspan="3">Language indicated on cut and individual speaker level</td></tr>
<tr><td colspan="3">Total speech seconds provided across regions (Insight B*)</td></tr>
<tr><td colspan="3">Speaker photo provided (Insight I*)</td></tr>
<tr><td colspan="2">Register indicated (Insight J*)</td><td>Register omitted (Insight J*)</td></tr>
</table>

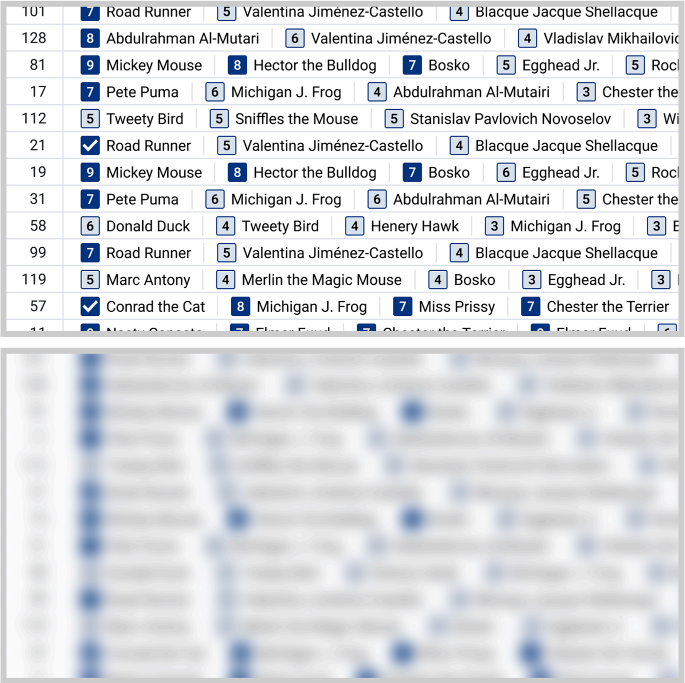
5.3. Depth of engagement in the prototypes. The development of the interface prototypes was informed by the depth of engagement framework, both in terms of how information was distributed among Tools 1 and 2, and how visual representations were leveraged within a given tool.

Table 10 provides an example of the depth of engagement framework’s influence on, and sensemaking of, visual representational relationships of the Square Digits scheme. There are three levels of depth indicating confidence: *notice* (binary differentiation), *read* (graded or interval differentiation), and *probe* (verbal descriptive differentiation).

Figure 4 further differentiates the *notice-read* distinction within the Square Digits scheme by simulating the first one or two fixations on a Tool 1 “scene,” in which the binary distinction of outlined vs. solid box is apparent. The blurring technique is a conventional means of approximating the visual perceptual information available in the initial fixations that serve to guide subsequent eye movements. It is not a simulation of perception — only of what information is processed perceptually.

Table 10. Three engagement depth levels in the Square Digits scheme.

Depth Level	Representation or Representational Facet	Information Layer Format and Semiotic Mode
notice	□ ■	**Binary:** higher vs. lower confidence Tonal variation
read	6 7	**Interval:** eight graded levels (2,3,4...9) Numeric-symbolic
probe	Fairly high confidence Moderate confidence	**Qualitative:** verbal descriptor Linguistic-symbolic



101	7 Road Runner	5 Valentina Jiménez-Castello	4 Blacque Jacque Shellacque		
128	8 Abdulrahman Al-Mutari	6 Valentina Jiménez-Castello	4 Vladislav Mikhailovic		
81	9 Mickey Mouse	8 Hector the Bulldog	7 Bosko	5 Egghead Jr.	5 Roc
17	7 Pete Puma	6 Michigan J. Frog	4 Abdulrahman Al-Mutairi	3 Chester the	
112	5 Tweety Bird	5 Sniffles the Mouse	5 Stanislav Pavlovich Novoselov	3 Wi	
21	✓ Road Runner	5 Valentina Jiménez-Castello	4 Blacque Jacque Shellacque		
19	9 Mickey Mouse	8 Hector the Bulldog	7 Bosko	6 Egghead Jr.	5 Roc
31	7 Pete Puma	6 Michigan J. Frog	6 Abdulrahman Al-Mutairi	5 Chester the	
58	6 Donald Duck	4 Tweety Bird	4 Henery Hawk	3 Michigan J. Frog	3 E
99	7 Road Runner	5 Valentina Jiménez-Castello	4 Blacque Jacque Shellacque		
119	5 Marc Antony	4 Merlin the Magic Mouse	4 Bosko	3 Egghead Jr.	3
57	✓ Conrad the Cat	8 Michigan J. Frog	7 Miss Prissy	7 Chester the Terrier	

Figure 4. The Square Digits scheme with a blur effect simulating what information is available in gist processing: the first one or two fixations reveal its internal *notice* layer.

The dual-coding of the *notice-read* layers in Square Digits is not redundant, as the *notice* layer is a classification of the more granular *read* layer. These layers are inseparable graphically, but they map onto the reading experience depth levels. The other schemes — Arc Gauge and Bar Fill — are also dual-coded, but their *notice-read* layers are redundant: the *notice*-layer color (green, yellow, red) corresponds directly with the *read*-layer outlines (gauge dial or fill line at 3/4, 1/2, 1/4).

5.4. Revisiting the research question. We now return to the research question, having described the interface prototypes and the visualization strategies they employ to indicate speaker model confidence:

> How can interactive visualizations of confidence indicators enable language analysts to more accurately interpret and efficiently act on speaker model outputs?

Other kinds of research questions can guide the user testing that is to follow this design investigation. As the above is a purely design-oriented research question, its answers are embedded in design exploration — they describe what has been demonstrated. (Section 6 extends the following list.)

1. Information that is not consistently needed by analysts can be placed — or *buried* — at greater "depths of engagement" that users can access — or *uncover* — on the occasions when the information becomes relevant and desirable.
2. Dual-coded representations can have internal engagement layers that align with the engagement levels that viewers traverse over thin slices of time, through user interaction. Surface layers can isolate actionable distinctions (e.g., binary: high enough confidence to open the media player [Tool 2], low enough confidence to ignore and scan to further rows [still in Tool 1]).
3. Confidence indicators can deviate from technology-relevant outputs (e.g., 101 levels: 0–100%) to provide knowledge worker–relevant outputs of utilitarian granularity (e.g., three levels: fair, good, excellent).
4. Visualization strategies for confidence indication can simplify confidence levels through symbolic representations (e.g., numbers in Square Digits), visual metaphors (e.g., tachometer metaphor in Arc Gauge), and visual analogies (e.g., volume in Bar Fill).
5. A confidence threshold user setting (as in Square Digits) can permit analysts to align visibility situationally with mission criticality through variable confidence indicators.

Confidence indicators can help users calibrate trust in AI in human-machine teams (Zhang et al., 2020), and our work with language analysts suggests that percentage-based confidence scores can be re-represented to improve trust calibration. Alternative forms of representation are available for speaker model confidence, some of which will be more meaningful to users, and some of which have affordances for user interaction that may further enhance trust calibration in human-machine teams.

6. Survey and Interview Insights

We conducted a survey of 16 language analysts at a formative phase of the project, featuring two UX systems and six visualization strategies. This resulted in **Insights A–L**, collected here:

A. **UX system:** The new features are preferred over the current systems by analysts — these "new features" are those cataloged in Tables 8 and 9. UX System A: 3/16 strongly agree, 7/16 agree; UX System B: 2/16 strongly agree, 8/16 agree.

B. **Speech seconds:** Indicating speech seconds in Tool 1 provides helpful information to analysts. Responses were split between Tool Tip (8/16) and Responsive Bar (6/16).

C. **Truth marking input:** Analysts prefer the binary speaker ID truth marking system (10/16). Some were interested in a five-point scale, but pointed out possible issues including confusion about how granular data might inform the model (5/16).

D. **Truth marking history (1):** Analysts prefer Total Up-Down Arrows (9/16) to Total Thumbs Up (4/16). Analysts felt that more information was better. However, the fact that thumbs down is not being used to train speaker models was not addressed in comments.

E. **Truth marking history (2):** Analysts participating in a separate live feedback session were confused about the term *truth marking history.* They had concerns around conflating truth marking with model updates.

F. **Truth marking averages:** Analysts prefer to exclude truth marking averages (7/16), though some favor inclusion (4/16).

G. **Alternative speakers:** Analysts prefer to be shown a list of alternative speakers — 13/16 supported in comparison to 2/16 opposed. Analysts also provided positive feedback on the list adjusting in response to thumbs-up and thumbs-down. They emphasized the need to edit speaker names directly.

H. **Visual threshold:** Analysts prefer being able to adjust the visual threshold to determine the lowest level of certainty that will be visually most noticeable, as is possible only in the Square Digits scheme — 11/16 supported, 2/16 opposed.

I. **Speaker photo:** Analysts prefer the inclusion of speaker photos — 12/16 supported, 1/16 opposed.

J. **Speaker register:** Analysts have mixed feelings regarding the inclusion of speaker register — 6/16 supported inclusion, 4/16 opposed. Six additional analysts did not vote but some of them shared negative feelings about register.

K. **Visualizations:** As implemented in the survey, analysts preferred these visualizations: Square Digits (3.6/5), Arc Gauge (3.3/5), Bar Fill (3.3/5), and Circle Check (2.8/5). However, the final three visualization strategies are modified schemes from those implemented in the survey — no data exists on improvements inherent to the refined versions.

L. **Color:** Analysts prefer the use of color as a secondary indicator for confidence level (i.e., analysts prefer dual-coding when possible).

An earlier survey and coordinated interviews resulted in **Insights M–U** (originally numbered 1–9 in documentation shared with LAS), with corresponding pain points. The following insights helped us understand how the existing systems are working for analysts — they do not refer to our interface prototypes and their features.

M. **Confidence scores (1):** Most analysts believe that confidence scores measure the probability that the identified speaker is their target. Analysts assume scores use logical, consistent measures. However, analyst experience demonstrates that confidence scores do not carry equal weight, leading to confusion.

   a. **Pain point:** Analysts are confused by the varying weights of confidence scores. These scores do not have equal weight because model health is not taken into account. The system does not, for example, differentiate between high-confidence but unverified models and well-trained, reliable models. (The remaining pain points are omitted here for space; refer to *XAISM-Pain-Points.pdf*.)

N. **Confidence scores (2):** Analysts use confidence scores as a threshold for action, but individual analyst thresholds vary based on their experience level, the criticality of the storyline, and other contextual factors. This leads to inconsistency in how these scores are applied.

O. **Confidence scores (3):** Analysts do not understand why the model applies a particular confidence score to a speaker — they do not understand how the model works. Analysts are unable to query the system to better understand the model's reasoning.

P. **Confidence scores (4):** Analysts cannot adjust confidence results based on the larger context of their investigation, leading to visual clutter and decreased relevance of results.

Q. **Confidence scores (5):** The speaker, language, and register scores are not of equal value to the analysts, yet they have equal visual presence.

R. **Model health:** Analysts cannot efficiently access useful model health data early in the query process (e.g., truth mark count, training data volume) to calibrate trust around the score.

S. **Truth marking (1):** Analysts often rely on truth marking to validate speaker IDs, but analyst truth marking behavior varies.

T. **Truth marking (2):** Analysts grow frustrated when their truth marking is not immediately reflected in model output. Confusion around this time lag reduces analyst trust in models and discourages truth marking.

U. **Truth marking (3):** Analysts' uneven truth marking behavior, combined with the unequal weight of thumbs up and thumbs down, skews the perception of model reliability.

7. Implementation Resources

The primary project deliverables for implementation are:

1. **Scenario videos:** a primary resource that details all proposed features, in many cases in alternate versions (as tracked in Tables 8 and 9).
2. **Confidence indicator icons:** individual files in SVG and EPS format for implementing each of the three selected visualization strategies — Square Digits, Arc Gauge, and Bar Fill.
3. **Specifications for developers:** visual markup identifying and explaining system features for the interface prototypes and visualization strategies — native in Figma and output as PDFs.
    a. XAISM-Dev-1-Color-Fonts-Icon-Alignment.pdf (the "XAISM" stem stands for Explainable AI Speaker Models)
    b. XAISM-Dev-2-Layouts.pdf
    c. XAISM-Dev-3-Components.pdf
    d. XAISM-Dev-4-Square-Digits-Walkthrough.pdf
    e. XAISM-Dev-5-Arc-Gauge-Walkthrough.pdf
    f. XAISM-Dev-6-Bar-Fill-Walkthrough.pdf

These resources will be compiled by LAS and/or on the high side.

We have also prepared various additional resources that may inform testing, implementation, or future development work. These resources are collected as PDFs:

1. **XAISM-Confidence-Survey.pdf:** results from a survey distributed to 19 language analysts to better understand their current experience interpreting confidence scores in the context of Tools 1 and 2.

2. **XAISM-User-Journey.pdf:** a user journey map used in problem definition.
3. **XAISM-Pain-Points.pdf:** data compiled from language analyst interviews and a language analyst experience survey.
4. **XAISM-Design-Survey.pdf:** results from a survey of 16 language analysts that A/B tested early confidence visualization strategies and distinct interactive features for Tools 1 and 2.
5. **XAISM-Insights.pdf:** compiled insights and feature recommendations from another survey ("UI Form Development Survey").
6. **XAISM-Accessibility.pdf:** accessibility evaluation. All icon elements passed level AA mandates for minimum contrast ratio according to Web Content Accessibility Guidelines (WCAG, 2.0). All icons also use color as a secondary indicator in concert with a primary indicator for individuals who cannot easily differentiate. More detailed notes are included in the PDF.

8. Recommendations

8.1. Proposed user testing. Our team recommends that next steps focus on A/B/C testing of the three proposed visualization schemes — Square Digits, Arc Gauge, and Bar Fill — as well as corresponding features to determine which options perform best. Primary testing should focus on the three proposed visualization strategies. We recommend that users (language analysts) experience these schemes within the user interface systems demonstrated in the three scenario videos — with one caveat. Our videos include some feature variation in Tool 2 (as indicated in Table 8). We included this feature variation in Tool 2 to capture rich insights gathered throughout this project. The caveat is that we recommend that primary testing eliminates this feature variation to isolate visualization strategy impacts (see Table 11). Once a visualization strategy preference emerges, Tool 2 feature variation should be A/B tested (see Table 12).

Table 11. Cross-tool visualization schemes for A/B/C testing.

A. Square Digits	B. Arc Gauge	C. Bar Fill
Square Digits visualization used	Arc Gauge visualization used	Bar Fill visualization used
Visual confidence threshold setting available	No visual confidence threshold setting available	

Table 12. Tool 2 feature variations for A/B testing.

UX Option A	UX Option B
Last person to truth mark indicated	No indication of last person to truth mark
Truth marking input: thumbs up/down. Click a second time on thumbs up/down to deselect.	Truth marking input: three-point scale. Click on the central point to deselect.
Register indicated	Register omitted

8.2. Future considerations. The following recommendations include features that analysts favored in user feedback but that are not currently operational in existing tools.

Speech seconds in Tool 1:

- Add individual speech seconds data. Include individual speech seconds with confidence descriptor language in a single pop-up.
- Test using hover to reveal the tooltip info (confidence level) and click to reveal the number of seconds for the individual speaker.
- Test responsive bar feature as a visual indicator of total speech seconds. This could be converted to individual speech seconds, data that analysts requested for Tool 1.
- Add a visual indicator of the number of total speakers. Tool 1 cannot currently distinguish how many total speakers are represented. This confuses analysts as they do not know whether the speakers indicated represent different individuals or different speaker ID options for the same individual.

Truth marking input in Tool 2:

- Convert currently specified three-point scale to a five-point scale, if and only if intermediate graded positive and negative indications could be used to train speaker models. For instance, a selection of *strongly agree* (5/5) might verify the speaker in the interface and provide heavily weighted input for model training, while a selection of *somewhat agree* (4/5) might not verify the speaker visually while still providing lightly weighted input for model training.

9. References

Burks, A. W. (1949). Icon, index, and symbol. *Philosophy and Phenomenological Research, 9*(4), 673–689. https://doi.org/10.2307/2103298

Knack, A., Carter, R. J., & Babuta, A. (2022). *Human-machine teaming in intelligence analysis: Requirements for developing trust in machine learning systems.* Centre for Emerging Technology and Security. https://cetas.turing.ac.uk/sites/default/files/2022-12/cetas_research_report_-_hmt_and_intelligence_analysis_vfinal.pdf

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Oliva, A. (2005). Chapter 41: Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251–256). Elsevier Academic Press. http://olivalab.mit.edu/Papers/Oliva04.pdf

Sadoski, M., & Paivio, A. (2001). *Imagery and text: A dual coding theory of reading and writing.* Lawrence Erlbaum.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). https://doi.org/10.1145/3351095.3372852