MasterWord®

How to Evaluate AI Solutions for Meaningful Language Access

With Use Case Evaluation Toolkit

Contents

DISCLAIMER OF LIABILITY

This white paper is intended for informational purposes only and is not a substitute for legal, professional, or regulatory advice. The content reflects research and analysis as of the publication date but may not address all regulations, relevant laws, or standards applicable to your specific organization or jurisdiction. Given the rapidly evolving nature of AI and language access technologies, readers should consider that some information may become outdated or require updates over time.

Readers are encouraged to verify all recommendations against best practices, current laws, and standards in their industry or jurisdiction. MasterWord Services, Inc. assumes no liability for decisions made based on this document and strongly advises readers to exercise independent judgment and seek expert input where necessary.

Copyright © 2025 MasterWord Services, Inc.

Evaluating AI Solutions Through the Lens of Meaningful Access

Integrating AI into language services requires a deliberate and ethical approach to ensure the technology enhances, rather than compromises, meaningful access. Agencies, organizations, and vendors must clearly disclose when AI is being used in interpreting or translation services, fostering trust and autonomy among end-users. Developers and providers of AI solutions must ensure that their systems are designed and maintained with the highest ethical standards. Buyers of services must also provide needed materials and invest in system training. Together, they must address any inaccuracies and biases, take responsibility for them and develop processes to fix them when they do occur.

Recommendations to Ensure Meaningful Language Access When Deploying AI

1. **Conduct comprehensive inventories of language access practices** to identify use cases where AI may deliver clear benefits.

   Evaluate current practices to identify areas where AI adds value, such as adding new touchpoints for language access, improving turnaround times, or managing high-volume tasks.

2. **Develop procurement and quality assurance guidelines** that incorporate industry standards from committees such as ASTM F43 and ISO/TC 37/SC 5.

   The National Technology Transfer and Advancement Act (NTTAA) of 1995 directs federal agencies to adopt voluntary consensus standards developed by industry, wherever practical, instead of creating government-unique standards. This policy promotes efficiency and cost savings while ensuring federal programs benefit from industry expertise.

3. **Establish a decision framework** to evaluate AI use based on complexity, context, and risk, using principles from:

   - Guidance on AI and Interpreting Services – Stakeholders Advocating for Fair and Ethical AI in Interpreting (SAFE AI) Task Force
     Outlines 4 ethical principles for the use of AI solutions for Interpreting
   - Automated Speech-to-Speech Interpreting – Six Evaluation Dimensions for Professional Deployments – CSA Research
     Provides 82 elements to investigate when selecting an AI tool
   - Guidance for Contracting AI-generated Interpreting (July 2024) – National Council on Interpreting in Health Care (NCIHC)
   - LEP.gov Language Access Plans – US Department of Justice

AI Interpreting Use Case Evaluation Worksheet

Not every situation is a good fit for AI. For example, in cases where someone is experiencing an active psychotic episode, a qualified human interpreter is essential because they bring the sensitivity and expertise needed for such interactions.

To make the most of your resources, focus on equipping your organization with tools that enhance language access while setting clear guidelines about when AI is appropriate within a specific setting. Carefully define use cases and ensure staff understand the boundaries to prevent misuse and make sure AI is applied effectively and ethically where it can provide real value.

*Use the following table to evaluate suitability of AI interpreting for your use case. After analyzing the factors impacting suitability assign a determination of a) **Suitable**, b) **Borderline**, or c) **Unsuitable** for each dimension.*

Evaluating Use Case Suitability

Use Case Dimensions

1. Category type: E.g., is the interaction appointment scheduling, a check-in at a registration desk, a legal hearing?
Describe the context of the interaction.

Factors Impacting Suitability:

- Criticality of the setting (e.g., routine vs. high-stakes interactions).
- Legal or regulatory requirements for language access.
- Sensitivity of the content (e.g., legal or medical implications).
- Complexity of information to be conveyed (e.g., simple instructions vs. nuanced dialogue).
- Availability of fallback options (e.g., escalation to human interpreters).
- Time sensitivity of the scenario (e.g., emergency situations).
- End-user familiarity and comfort with AI tools.

What policy might you put in place when AI may not be appropriate?
***Example:** Establish a **default human interpreter policy** for high-stakes or sensitive scenarios, such as legal hearings, medical diagnoses, or any situation requiring trauma-informed communication. Require human interpreters for all cases with legal or regulatory compliance needs and include a clear escalation protocol to switch from AI to a human interpreter when AI is deemed insufficient.*

☐ **Suitable** ☐ **Borderline** ☐ **Unsuitable**

2. Interaction type: E.g., might loud ambient noise or interaction with an individual with cognitive impairment or speech be expected?

Describe any environmental factors that may impact the effectiveness of an AI solution.

Factors Impacting Suitability:

- Presence of background noise or poor acoustics.
- Variability in speaker accents, clarity, or speeds.
- Potential for cognitive or speech impairments affecting communication.
- Reliability of the technology infrastructure (e.g., microphones, internet connection).
- Availability of noise-canceling or voice isolation tools.
- Use of additional media inputs (e.g., on-screen text or visual aids).

What policy might you put in place when AI may not be appropriate?

Example: *Implement an* ***environmental suitability checklist*** *to assess interaction conditions before deploying AI. If loud ambient noise, cognitive/speech impairments, or poor audio quality are identified, mandate the use of human interpreters. Provide on-site troubleshooting options to quickly address technical issues or escalate to a human interpreter.*

☐ **Suitable** ☐ **Borderline** ☐ **Unsuitable**

3. Interactivity type: E.g., how many speakers and what degree of overlapping dialogue might be expected?

Outline the expected level of interactivity, detailing the number of participants, potential for overlapping dialogue, and whether the conversation requires back-and-forth exchanges or is primarily unidirectional.

Factors Impacting Suitability:

- Number of speakers and their interaction dynamics.
- Frequency and intensity of overlapping dialogue.
- Degree of conversational flow complexity (e.g., structured vs. freeform).
- Importance of turn-taking and speaker identification.
- Expected duration of back-and-forth exchanges.
- Degree of participant engagement required (e.g., passive listening vs. active Q&A).
- Availability of tools to manage interactivity, such as moderated turn-taking systems.

What policy might you put in place when AI may not be appropriate?

Example: *Adopt a* ***complex interaction policy*** *that requires human interpreters for discussions with additional complexity such as conversations with unpredictable turn-taking or overlapping dialogue. Define a maximum threshold for participants and interactivity complexity beyond which AI cannot be deployed. Incorporate tools for monitoring real-time interaction dynamics to ensure smooth handoffs to human interpreters when necessary.*

☐ **Suitable** ☐ **Borderline** ☐ **Unsuitable**

4. Interpreter role: E.g., is the interpreter expected only to interpret words or also to engage in advocacy or culture mediation?

Specify if meeting the meaningful access threshold will require more than simple linguistic transfer to support nuanced communication.

Factors Impacting Suitability:

- Need for cultural sensitivity and contextual understanding.
- Requirements for trauma-informed communication skills.
- Role of advocacy or mediation in resolving misunderstandings.
- Sensitivity of the conversation topic (e.g., mental health, personal trauma).
- Need for interpreters to convey emotional tone or non-verbal cues.

What policy might you put in place when AI may not be appropriate?

Example: *Enforce a* ***context-specific interpreter role policy*** *to require human interpreters in cases needing advocacy, cultural mediation, or emotional intelligence. AI should only be used for straightforward linguistic translation, with mandatory fallback to trained human interpreters for conversations involving cultural nuances or trauma-sensitive topics.*

☐ **Suitable** ☐ **Borderline** ☐ **Unsuitable**

5. Language Pair(s): E.g., K'iche'-English, Mandarin-English, or Spanish-English

Request vendors to provide accuracy ratings for specific language pairs based on their internal testing results. For instance, a solution might deliver a 95% accuracy rate for Spanish-English for certain interactions but only 80% for Mandarin-English.

Factors Impacting Suitability:

- Reported accuracy ratings for the specific dialects, language pair, and regional accent.
- Level of customization or training available for AI systems in the given languages.
- Availability of well-trained human interpreters for low-performing pairs in AI.
- Frequency of use for the language pair in your organization.

Evaluate whether the reported accuracy aligns with your quality requirements and prevent the use of language combinations that fall below acceptable performance levels in non-testing scenarios.

What policy might you put in place when AI may not be appropriate?

Example: *Implement a* ***language accuracy threshold policy*** *that restricts AI deployment to language pairs with accuracy rates where your own testing by language and use case consistently reaches a defined percentage (e.g., 95%). Require human interpreters for languages of limited diffusion, dialects with insufficient AI training, or scenarios where language pair accuracy could result in miscommunication risks.*

☐ **Suitable** ☐ **Borderline** ☐ **Unsuitable**

6. Accuracy: E.g., what definition of accuracy will meet your organization's meaningful access threshold?
Define acceptable accuracy based on your organization's testing for the exact type of use case and languages. Note that results will vary by language.

Factors Impacting Suitability:
- Risk of miscommunication for high-stakes interactions.
- Impact of misinterpretation on decision-making outcomes.
- Impact of errors on the audience or service recipients.
- Experience required for domain-specific interactions (e.g., legal or medical expertise).

What policy might you put in place when AI may not be appropriate?
Example: *Implement a* ***language accuracy threshold policy*** *described in dimension 5.*

☐ **Suitable** ☐ **Borderline** ☐ **Unsuitable**

7. Security & Privacy: E.g., will legal case details, patient data, or personal financial information be involved?
Determine whether the use case involves handling confidential or sensitive information.

Factors Impacting Suitability:
- Sensitivity of the data being communicated (e.g., financial, medical, or personal).
- AI solution compliance with data privacy regulations (e.g., GDPR, HIPAA).
- Risk of data breaches or misuse of recordings or transcripts.
- Transparency of vendor data retention and deletion policies.
- Ability to limit or restrict AI training based on sessions.
- Need for confidentiality assurances from participants.
- Risk assessment for sharing sensitive data via automated systems.
- Impact of data mismanagement on organizational reputation or liability.

Consider SOC 2 (System and Organization Controls 2) compliance to assess systems and processes.

What policy might you put in place when AI may not be appropriate?
Example: *Adopt a* ***data sensitivity policy*** *that prohibits the use of AI for interactions involving confidential or sensitive data unless the solution demonstrates full compliance with privacy regulations (e.g., GDPR, HIPAA). For sensitive settings like legal proceedings or medical consultations, require human interpreters and enforce strict data handling protocols to prevent breaches or misuse.*

☐ **Suitable** ☐ **Borderline** ☐ **Unsuitable**

Determining Best Use Case Solution

All dimensions SUITABLE	***You can experiment with AI interpreting or a hybrid solution***
One or more dimensions BORDERLINE	***You can experiment with a hybrid solution if you have a strong process to integrate qualified human interpreters*** **Risk Mitigation for Borderline Cases:** • **Hybrid Solutions:** Consider AI with human oversight or escalation to a qualified human interpreter for critical moments or for issues AI cannot resolve. • **Pretesting:** Run simulations to test AI effectiveness in these contexts before committing to deployment.
One or more dimensions UNSUITABLE	***Focus on working with qualified human interpreters***

Note: Consider the options laid out in this CSA Research table for fit-for-purpose solution (Page 11).


RULES
GOVERNANCE
COMPLIANCE
LAWS
STANDARDS

Considerations for Piloting AI Interpreting

1. **Validate Vendor Claims**
   - Evaluate performance metrics including error rates, especially for less-supported or low-resource languages.
   - Evaluate the potential for bias in output.
   - Insist on transparency in vendor documentation regarding accuracy, privacy, and security.
2. **Establish Guidelines and Controls**
   - Put policies in place to prevent the use of self-procured apps you did not validate for use.
   - Define clear rules for when and how AI interpreting can be used.
   - Equip frontline staff with a decision-making framework and escalation protocols to human interpreters.
   - Plan for device usage — placement, storage, charging, theft prevention, etc.
3. **Train and Prepare Users**
   - Work with professional-grade AI tools and invest in training to your specific needs.
   - Provide comprehensive training on AI interpreting tools, their limitations, and appropriate use cases.
   - Implement best practices for audio management.
   - Train participants on how to optimize results — good microphone, way of speaking, etc.
   - Address potential operational issues such as internet connectivity and hardware compatibility.
4. **Ensure Quality and Transparency**
   - Include human oversight mechanisms to monitor and intervene in high-stakes or complex cases.
   - Be transparent with stakeholders about the use and performance of AI tools and manage expectations.
5. **Monitor and Scale**
   - Continuously assess tool performance and refine usage strategies based on data from pilot programs.
   - Collect user feedback to identify persistent issues and determine when to expand AI use or adjust processes.

Key Metrics for Pilot Evaluation

Track these metrics during the pilot phase to assess the feasibility of scaling AI interpreting while identifying areas for improvement. Compare AI performance against the outputs of the **average interpreter your organization typically uses**.

1. **Latency:** Measure the time delay between spoken input and AI-generated output. Ensure latency remains acceptable for the use case, particularly in real-time interactions where delays can disrupt conversational flow.
2. **Voice Recognition Accuracy:** Evaluate the AI's ability to accurately transcribe spoken language, particularly in challenging conditions with accents, fast speech, or noisy environments.
3. **Subtitling Accuracy:** Use the subtitles (if available) to assess the accuracy of the translation portion of the AI process.
4. **Quality of Synthetic Voices:** Examine the intelligibility, naturalness, and tone of AI-generated voices. Synthetic voices should enhance communication and maintain the intended emotional or tonal nuances, aligning with user expectations.
5. **Error Impact Analysis:** Document and analyze negative outcomes caused by AI errors, such as delays, miscommunications, or safety risks. Use these insights to understand the limitations of the AI system and refine its application boundaries.

Source: *Automated Speech-to-Speech Interpreting, CSA Research*

Key Sources

- **Guidance on AI and Interpreting Services** – Interpreting SAFE AI Task Force
- **Automated Speech-to-Speech Interpreting – Six Evaluation Dimensions for Professional Deployments** – CSA Research
- **Guidance for Contracting AI-generated Interpreting (July 2024)** – National Council on Interpreting in Health Care (NCIHC)
- **LEP.gov Language Access Plans**
- **EU Artificial Intelligence Act**
- **Biden Administration Executive Order on AI**

MasterWord: Your Trusted Partner in AI-Enhanced Language Services

As the global landscape of language services evolves, MasterWord remains dedicated to providing impactful, innovative, and secure solutions tailored to your organization's needs. Contact us today to explore how our AI-enhanced, human-centered approach can help you achieve your language access goals.

LET'S ADVANCE LANGUAGE ACCESS TOGETHER

MasterWord®

Email: **salesgroup@masterword.com** | Phone: **1.281.589.0810**

www.masterword.com

Sponsor and participant in the SAFE AI Task Force, MasterWord is championing progress in ethical AI integration, prioritizing accountability, transparency, and user trust.

Connecting People Across Language and Culture®