Perceptyx AI: Model Governance & Overview

Driving Employee Experience through Ethical, Secure, and Actionable AI.

This document outlines key information about the AI models utilized to power Perceptyx's products. Our data science team customizes commercially available LLMs from reputable providers to align with the specific requirements of our offerings.

Our AI Model Portfolio

We utilize four distinct types of AI to ensure accuracy, privacy, and meaningful insights:

- **Discriminative AI:** Our core offering leverages proprietary AI Models that annotate employee comments with emotion, intent, noteworthiness, sentiment, and themes.
- **Generative AI (Opt-in Only):** Optional features that are powered by Open AI and Google.
- **Privacy AI:** Automated Named Entity Recognition (NER) to detect and mask names before processing.
- **Rule-Based AI:** Expert-defined logic for consistent, automated decision-making based on human knowledge.

Intended Use & Access

Primary Users: Central business teams, HR Admins, and Senior Leaders.
Intended Use: Our AI solutions are designed to deliver actionable employee feedback insights, strengthen listening strategies, facilitate individual coaching, and offer conversational learning tools to ensure lasting professional growth.
Access Controls: All access to AI features is role-based and can be controlled by the customer. Customers maintain complete control over the activation of Generative AI features through a strictly opt-in framework enabling organizations to leverage AI capabilities based on their own internal compliance requirements and comfort levels. These tools are designed as optional enhancements and are not required for the core platform to operate.
Guardrails: Guardrails are in place to prevent misuse and promote relevance. The AI draws from a curated library to provide responses and refuses to answer questions about sensitive topics.
Human-in-the-loop: Designed to be intuitive; users can pre-filter data and use curated prompts to ensure relevance.

Use Case	Description
Employee Listening	Enhancement through tools like Comment Copilot, Narrative Analysis, Conversational Listening, AI Coach, and Intelligent Nudges.

Use Case	Description
Actionable Insights	Providing insights from feedback using sentiment, theme, intent, and emotion modeling.
Personalized Support	Supporting coaching and behavioral change through AI-driven recommendations.
Organizational Analysis	Analyzing comment data from various sources to identify strengths and areas for improvement.
Validated Learning	AI Agent that adapts traditional corporate training into interactive, conversational experiences to validate comprehension.

Non-Intended Use: These tools and features are not intended to impact individuals outside the user organization or their rights and are not used to create content or make decisions or recommendations about individuals within the organization.

Privacy, Ethics & Bias Mitigation

Perceptyx adheres to a "Privacy-by-Design" philosophy. Our teams engage in cross-functional collaboration to develop policies and practices regarding the use and development of AI.

No Customer Data Used for Training: Neither Perceptyx nor our third-party LLM providers use customer data to train or fine-tune AI models.
Bias Reduction: Our Data Science team takes measures to minimize bias during AI model training. Our models are trained on augmented datasets to help reduce bias and increase robustness.
Anonymization: Respondent anonymity is protected through aggregation. Individual survey responses are not attributed to a named employee and reports are only generated once a sufficient number of responses are received. Strict minimum reporting thresholds are in place to prevent individual identification.

Training Data

Discriminative AI: We use separate proprietary and anonymized datasets to train our AI models. Training data are aggregated from a combination of open-source datasets and pre-cleaned comment datasets bootstrapped in-house.
Generative AI: In-house experts refine responses and engineer prompts based on workforce industry best practices.

Performance Factors

Variations in input training data, model parameterization, training regimen, and objective function. Demographic imbalances are addressed via pre-processing data augmentations (e.g., pronoun replacement) and in-processing procedures (e.g., over/under sampling) to measure and reduce bias.

Quality & Reliability

We treat AI performance as a rigorous engineering discipline:

Testing and Validation: Our AI models are evaluated on a separate validation set to promote accuracy and generalizability. Performance is measured using various metrics, including precision, recall, F1 score, accuracy, and ROC/AUROC curves.
Monitoring: We run internal quality tests to monitor accuracy, improve system performance, and refine responses over time.
Hallucination Control: While hallucinations and inaccuracies cannot be entirely eliminated, we take steps to minimize AI hallucinations through structured guidance, drawing responses from our curated sources, and state-of-the-art model selection.

Responsible AI Principles

Principle	Our Commitment
Accountability	Commitment to responsible AI practices that prioritize data privacy and security. Perceptyx is ISO 27001 certified and undergoes annual SOC 2 Type II audits and third-party penetration testing. We are actively pursuing ISO 42001 certification which is expected by the end of December 2026.
Security and Safety	All AI systems are tested and validated against separate, held-out datasets to demonstrate effectiveness. Metrics such as precision, recall, F1 score, accuracy and ROC/AUROC curves are reviewed to promote accuracy and generalizability. Our models undergo regular audits and the code is regularly reviewed statically, dynamically, and with external pentesters. Monitoring of product model metrics include interference output distributions, model drift metrics, and model confidence metrics to ensure ongoing reliability and security.

Privacy	Our automated NER model is applied before data is processed to detect and mask names in comments. Respondent privacy is protected through aggregation of results in reporting.
Fairness	Our models are trained on augmented datasets with artificial noise injected to help reduce bias and increase robustness. During model training, steps are taken to minimize bias and maximize fairness.
Transparency and Explainability	The Perceptyx Data Science team is able to trace and explain results of all Perceptyx AI systems. Inputs and outputs to AI systems are logged, along with other model performance metrics, allowing for transparency within the AI system. AI-assisted and AI-generated content is labeled within the platform. Outcome challenges are possible in the backend of the application.
Robustness	Models trained on augmented datasets with artificial noise to increase robustness.