

# Responsible AI

Ruth Yakubu  
Cloud Advocate  
[@RuthieYakubu](https://twitter.com/RuthieYakubu)

# Agenda

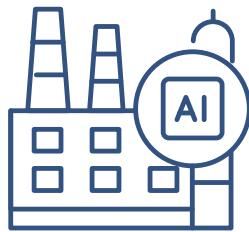
- ❑ Why Responsible AI?
- ❑ Azure Responsible AI dashboard
- ❑ OpenAI Content Safety
- ❑ Microsoft 365 for Inclusion



# The world today...



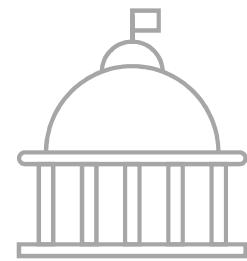
AI **innovation** is occurring at a rapid pace



Companies are accelerating **adoption** of AI



Societal expectations are **evolving**



Governments are **regulating** AI in response

# Societal expectations are increasing



**The New**

# A.I. Is M Language. Trust Wh

OpenAI's GPT-3 and o  
write original prose with  
a development that  
implications

## Can Administrators Ensure the Ethical Use of AI in K–12 Education?

School and IT leaders should consider the questions before investing in artificial intelligence and machine learning technologies for the classroom.

The New York Times

# A.I. Is Mastering Language. Should We Trust What It Says?

By Farhad Manjoo

ARTIFICIAL INTELLIGENCE IS BEING USED TO LEARN LANGUAGE. SHOULD WE TRUST WHAT IT SAYS?

It's been a remarkable year for AI. In January, it beat the world champion at Go. In March, it beat the best player at Dota 2. In May, it beat the best poker player in the world. In June, it beat the best Jeopardy! player. In July, it beat the best player at StarCraft II. In August, it beat the best player at Counter-Strike: Global Offensive. In September, it beat the best player at Super Mario Bros. And in October, it beat the best player at League of Legends.

It's also been a remarkable year for AI research. In January, researchers at Google DeepMind announced that they had created a system that could learn to play any video game. In March, researchers at Facebook AI Research announced that they had created a system that could learn to play any game in a certain class. In May, researchers at DeepMind announced that they had created a system that could learn to play any game in a certain class. In June, researchers at DeepMind announced that they had created a system that could learn to play any game in a certain class. In July, researchers at DeepMind announced that they had created a system that could learn to play any game in a certain class. In August, researchers at DeepMind announced that they had created a system that could learn to play any game in a certain class. In September, researchers at DeepMind announced that they had created a system that could learn to play any game in a certain class. In October, researchers at DeepMind announced that they had created a system that could learn to play any game in a certain class.

But there's one area where AI has not yet made much progress: language. In fact, it's still pretty bad at it. It can't understand what you're saying, and it can't say what you're thinking. It can't even understand what you're writing. It's like a baby who can't speak yet.

That's why I'm excited about the work that's being done in this field. I think it's going to be a breakthrough year for AI language research. And I think it's going to be a breakthrough year for AI language research.

The collage consists of three main parts. The top right is a digital interface with a grid of colored nodes (blue, green, red) and various icons like arrows and plus signs. The bottom right is a close-up photograph of a green printed circuit board (PCB) with several purple cylindrical components. The bottom left is a white document page with black text and a small blue logo in the top left corner.

# How Do Patients Feel About Using AI in Health Care?

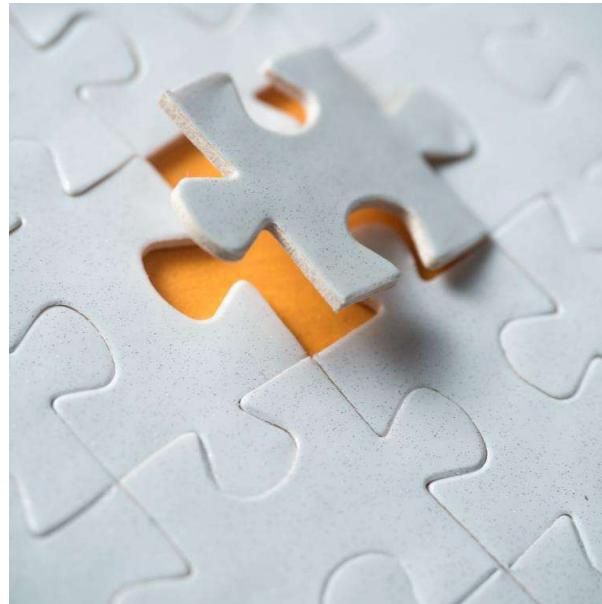
## **Deepfake Voice Used to Steal Over \$240,000 in AI-Powered Heist**

The robots are stealing money now.

# Why Responsible AI?



AI systems continue to miss expectations.



New AI breakthroughs expose new challenges.



Government regulation on AI imminent.

# Learn Microsoft's AI principles



Fairness



Reliability  
& Safety



Privacy &  
Security



Inclusiveness



Transparency



Accountability

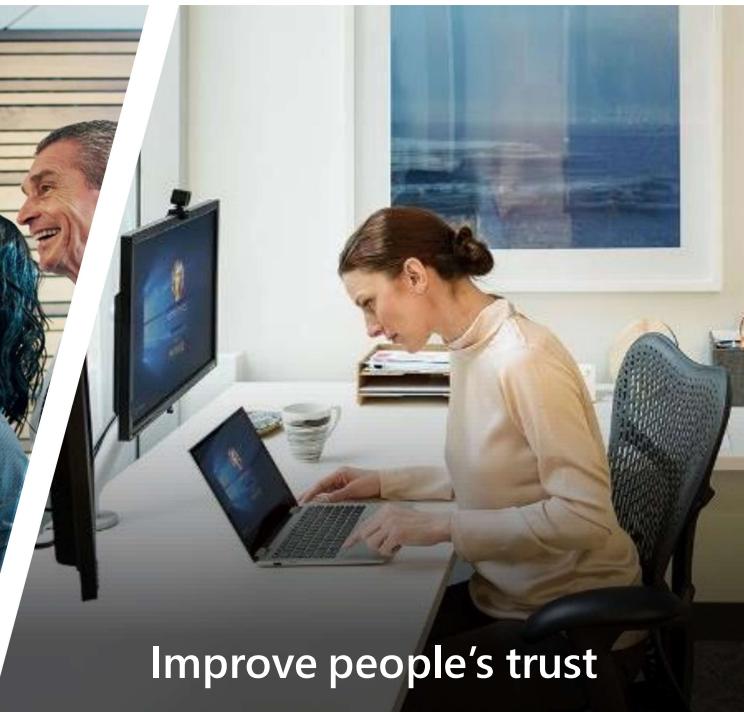
# What does this mean to you?



Machine Learning tools

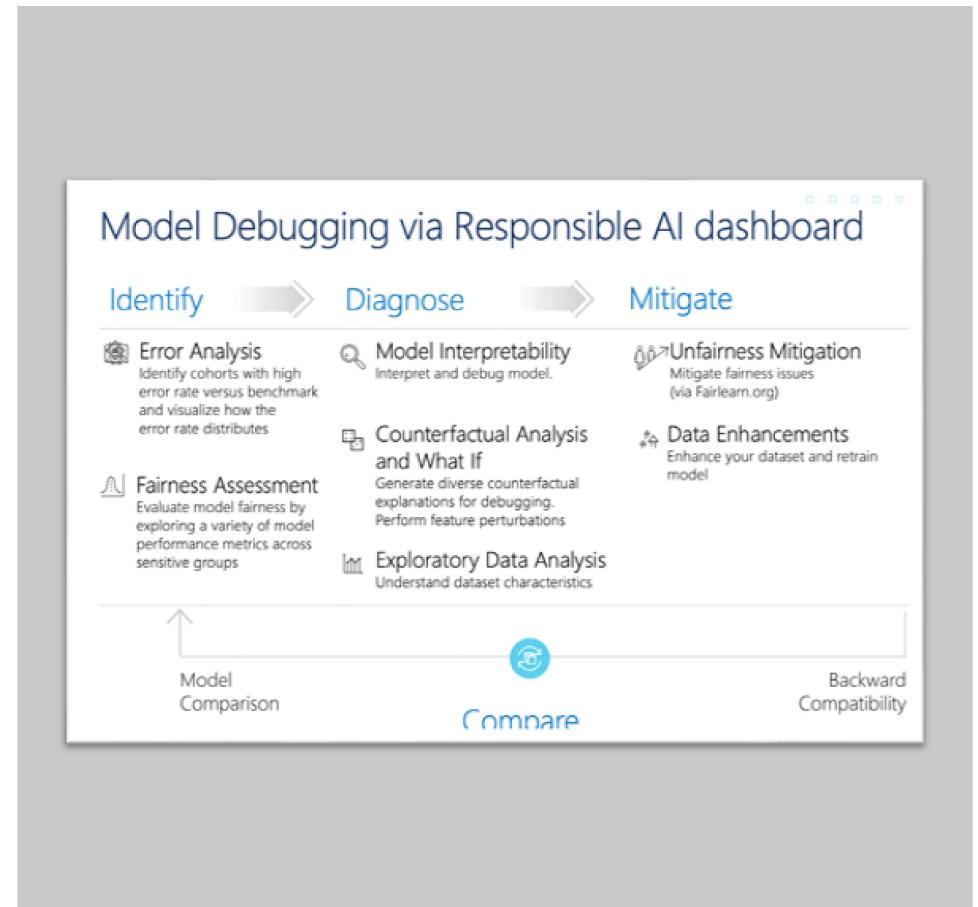
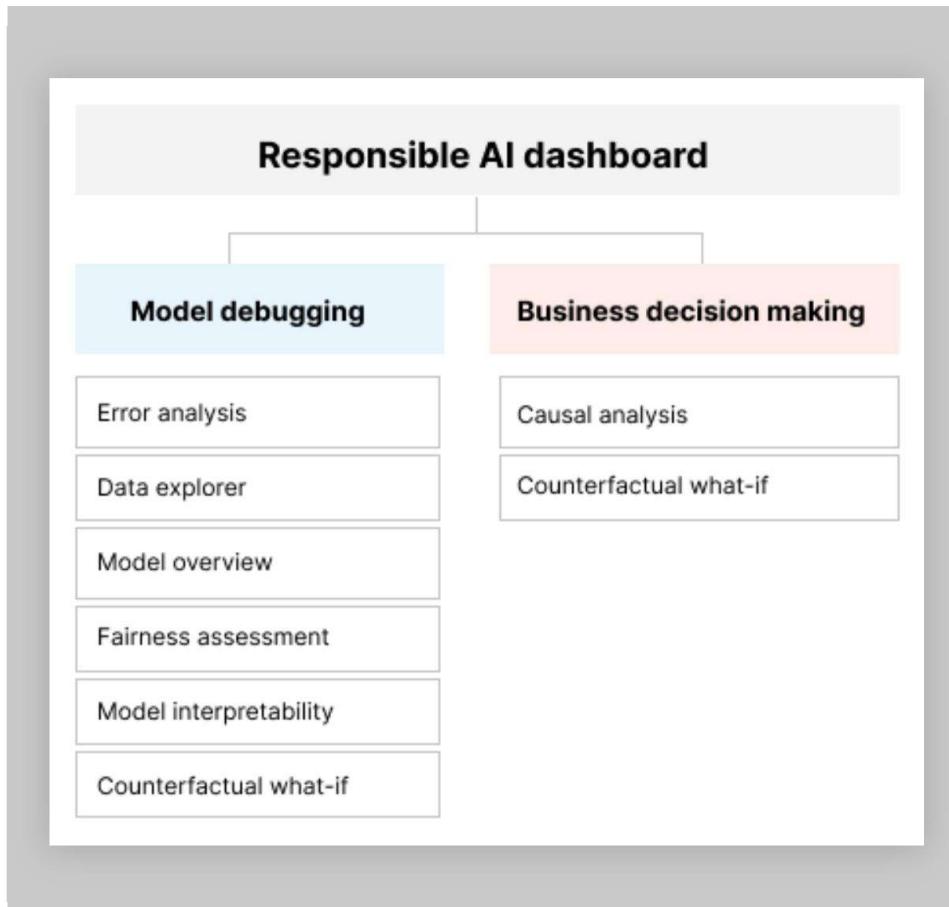


Enable decision-makers



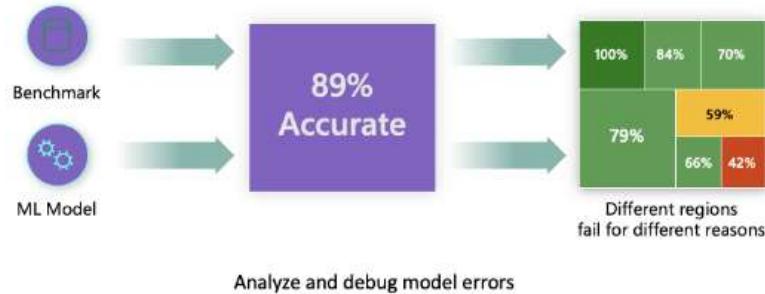
Improve people's trust

# Model debugging with Responsible AI dashboard

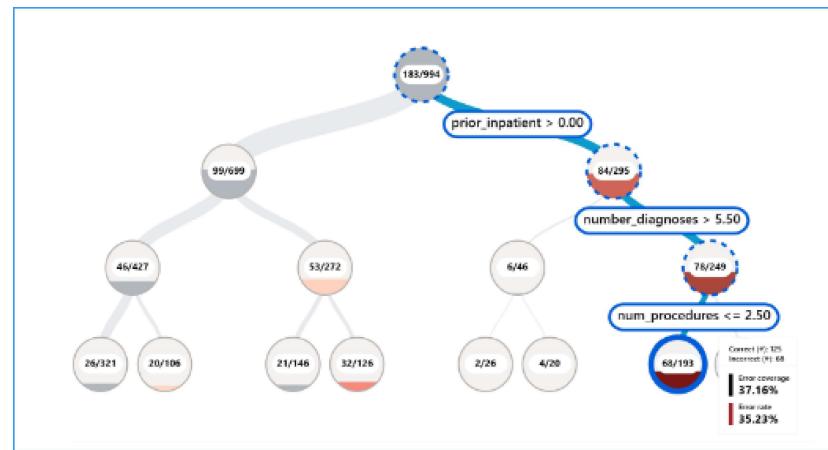


# Gaps in model error identification

## Error distribution



## Error Analysis

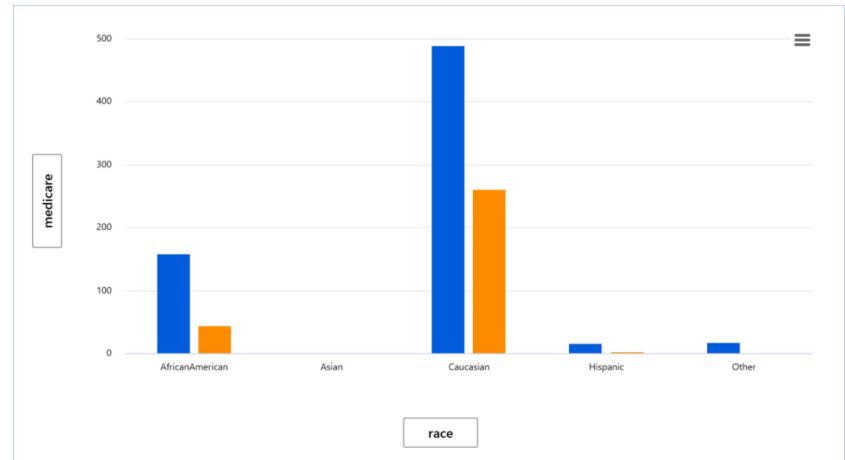


# Gaps in finding data bias

## DATA Representation

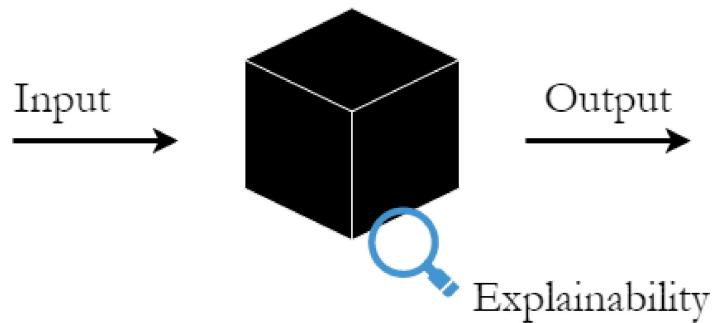


## Data Analysis

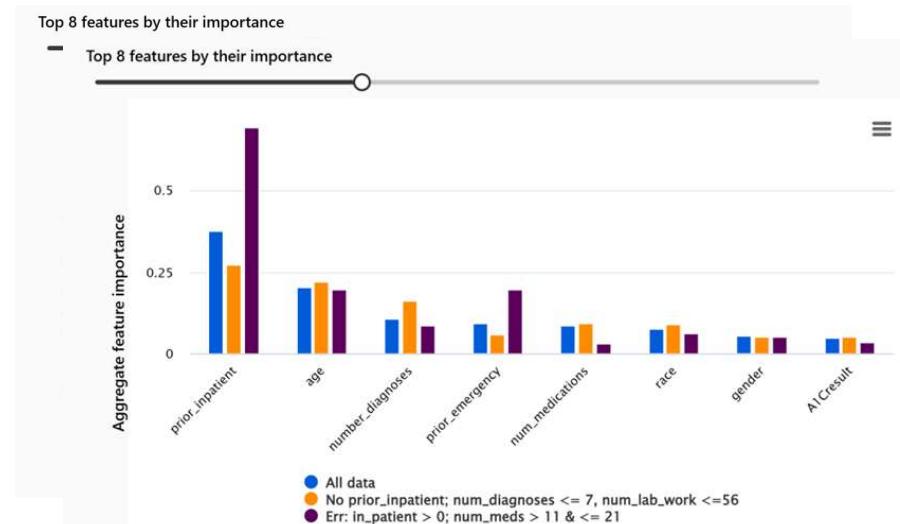


# Gaps in model explainability

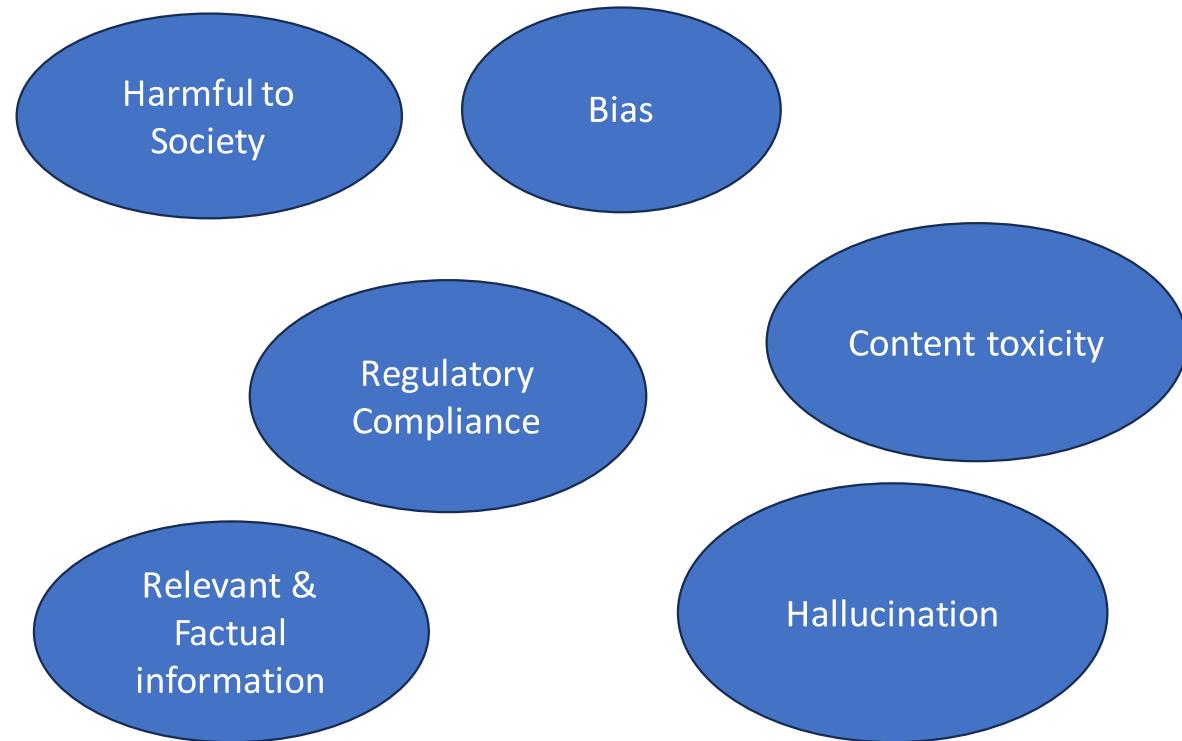
## Blackbox



## Feature Importance



# Open AI Challenges to consider



# Azure Content Safety

## Content Filters

The Azure AI Content Safety service works alongside core models by running both the prompt and completion to detect and prevent the output of harmful content.

Supported languages: English, German, Japanese, Spanish, French, Italian, Portuguese, and Chinese

 Classifies harmful content into four categories via Azure OpenAI API response

Hate

Sexual

Violence

Self-harm

 Returns a severity level score for each category from 0 to 6

2

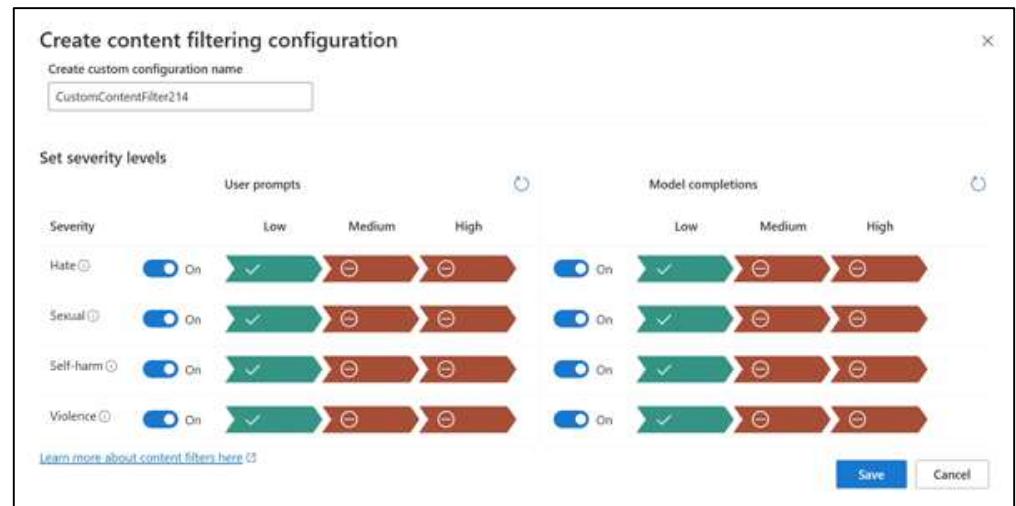
0

4

6

# Configurable Azure OpenAI Content Filters

Severity	Config for prompts	Config for completions	Description
Low, Medium, High	Yes	Yes	Strictest filtering configuration. Content detected at severity levels low, medium and high is filtered.
Medium, High	Yes	Yes	Default setting. Content detected at severity level low passes the filters, content at medium and high is filtered.
High	No	No	Content detected at severity levels low and medium passes the content filters. Only content at severity level high is filtered.



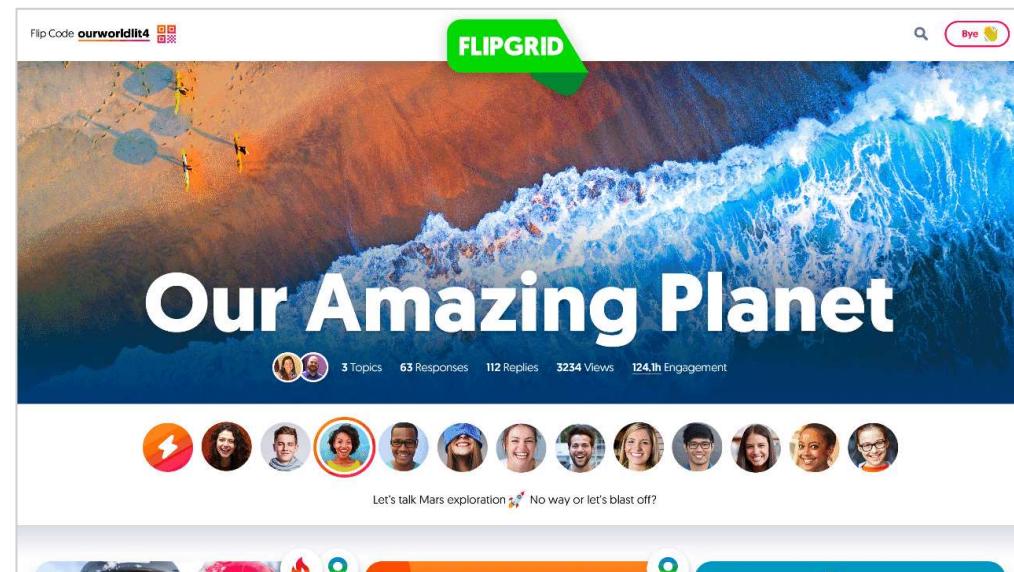
# M365 for Education

- Inclusively designed to help people of all abilities with reading.
- Dyslexia, ADHD, emerging readers, non-native speakers, visual impairments
- Accessibility and learning tools, including Immersive Reader, translation, transcription, speech-to-text, and more
- Already used in Word, OneNote, Teams, Edge, PPT, SharePoint, Outlook, Forms, Flipgrid, Minecraft EDU and Lens.



# Immersive Reader Capabilities

- Read Aloud and word/line highlighting (50 languages)
- Voice speed
- Reduce visual crowding
- Page colors tuned for accessibility
- Break words into syllables
- Highlight Parts of Speech (noun/verb/adj/adv)
- Line Focus
- Picture Dictionary
- Word or Document Translation (110 languages)



*Example of the Immersive Reader in [Flipgrid](#)*



# Automation, AI, and analytics

## Education Insights

Educators and educational leaders have access to data and insights for individual students, multiple classrooms, and across the entire institution—to help improve learning outcomes

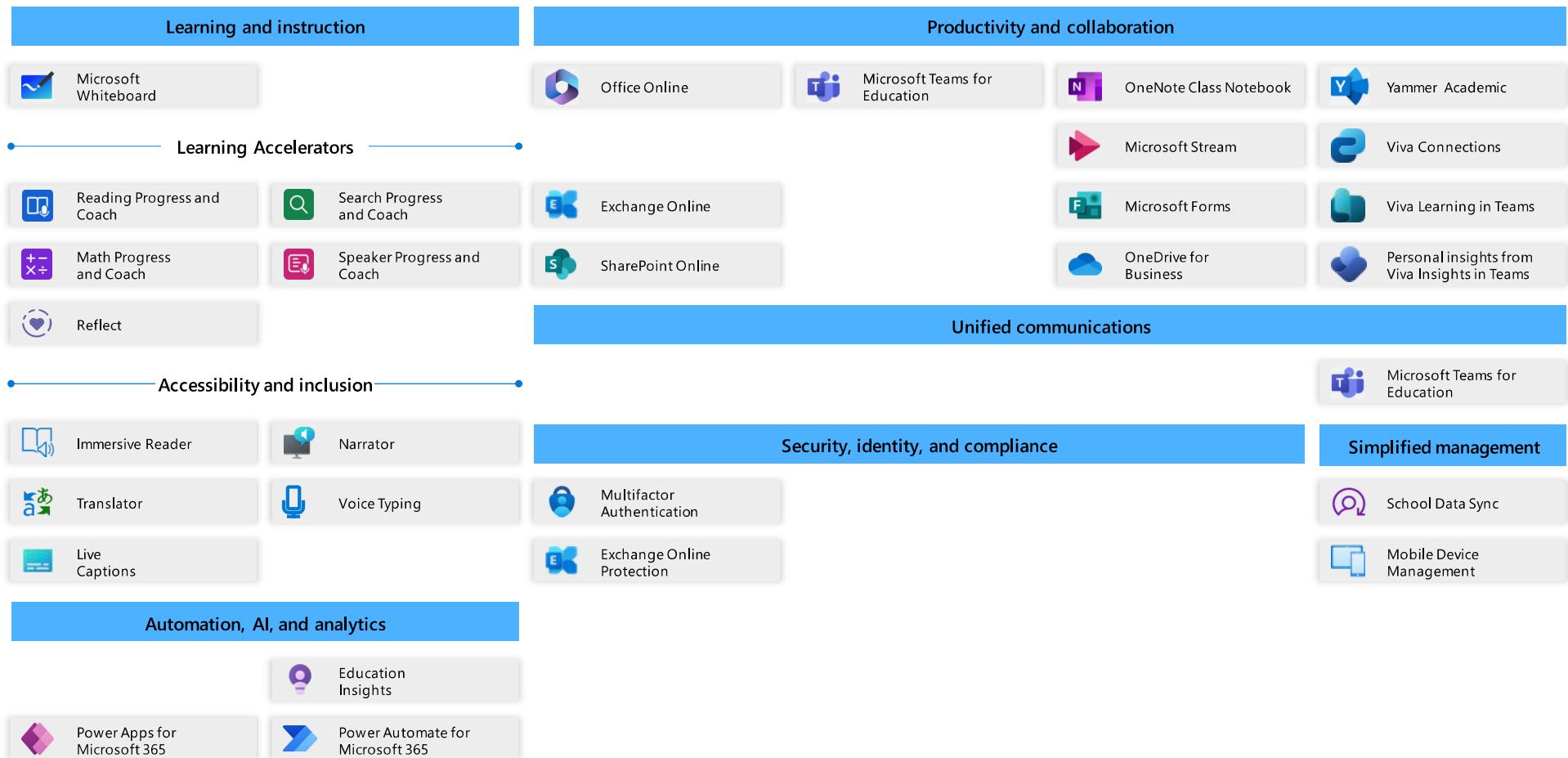
The Microsoft Teams Insights dashboard provides a comprehensive overview of student activity and performance. Key metrics displayed include:

- Digital engagement this week:** 3 inactive students (2 fewer than last week), 47 new communications (10 more than last week).
- Activity:** Alex Wilber wasn't active last week.
- Studying:** 40% of students work on assignments late at night.
- Communication:** 29 students participated in a Shakespeare plays discussion.
- Assignments this month:** 7 missed submissions (3 fewer than last month), 89 average grade (10 points higher than last month).
- Time management:** Ella Taylor hasn't started Comedy of Errors yet.
- Grades trend:** The average grade for Write a poem is relatively low, at 66 points.
- Workload:** Eldon Roque submitted 6 out of 8 assignments late.

Education  
Insights



# Microsoft 365 for Education





# Resources

- Microsoft Learn Educator: <https://aka.ms/Microsoft-Education-Center>
- Azure Content Safety: <https://aka.ms/Edu-ChatGPT-Content-Safety>
- Responsible AI dashboard: <https://aka.ms/Edu-ResponsibleAI-dashboard>
- Pre-built AI: <https://aka.ms/Edu-Prebuild-AI-Services>

**THANK YOU**