# AI and Language Processing Predictions for 2025

This document was automatically generated
using the Language Studio Business AI Feature

*You can design your own templates in minutes and generate a wide range of meeting and presentation extracts with the exact information that you want included.*

## Table of Contents

# Summary of AI Predictions Presentation

These predictions underscore the transformative potential of AI across industries, emphasizing its growing affordability, creative applications, and ability to drive compliance and operational efficiency. Businesses should focus on modular frameworks, real-time applications, and balancing automation with human oversight to unlock the full potential of AI in 2025.

## Additional Resources

Omniscien has provided the following supporting materials to complement the insights shared during the webinar:

1. **Detailed Blog Posts**: For each prediction, there is an in-depth blog post offering supporting analysis, actionable advice, and key insights.
2. **Webinar Replay**: Watch the full webinar recording to revisit the discussion.
3. **Slides from the Presentation**: Access the presentation slides for a visual summary of key points.
4. **Automatically Generated AI Analysis**: This document offers a comprehensive AI-generated analysis of the webinar.
5. **UFA Video Case Study**: Learn about the groundbreaking no-human-in-the-loop video localization workflow.
6. **Business AI Overview Video**: Explore how AI can transform business operations and enhance efficiency.
7. **Conversational AI Demonstration**: See real-time conversational AI capabilities in action.

All resources are available at: https://omniscien.com/blog/predictions/

# Webinar Overview

**Title and Theme**:
**"AI Predictions for 2025: Shaping the Future of Business and Language Processing"**
This webinar explored the transformative impact of AI on industries, focusing on trends in everyday business operations, modular AI frameworks, real-time translation, and language processing. The session highlighted key predictions for 2025, offering actionable insights for businesses navigating the rapidly evolving AI landscape.

**Date, Time, and Duration**:
The webinar took place on **January 19, 2025**, running for **90 minutes**, including a comprehensive Q&A session.

**Purpose**:
The webinar aimed to provide attendees with a forward-looking view of AI's role in shaping business strategies, creative processes, and technological infrastructure. It emphasized practical applications, innovations, and compliance considerations to help businesses stay competitive and informed in a rapidly advancing field.

**Audience**:
The webinar attracted a diverse audience of professionals, including business leaders, technology specialists, language service providers, and researchers. Over **500 participants** from industries such as technology, healthcare, finance, and government joined to gain insights into leveraging AI for innovation and operational efficiency.

**Key Takeaways**

1. **AI's Integration Into Everyday Business**:
   AI has moved beyond experimentation to become a crucial tool for improving efficiency, customer satisfaction, and innovation. Actionable advice included prioritizing AI applications with measurable ROI, adopting modular frameworks, and investing in employee training to integrate AI effectively.
2. **Affordability and Democratization of AI**:
   Cost reductions in AI models, driven by smaller, more efficient architectures and modular solutions, are making AI accessible to businesses of all sizes. Open-source models and on-demand services such as AWS Bedrock provide affordable options for scaling AI use.
3. **The Rise of AI Agents and Modular Frameworks**:
   2025 will see the widespread adoption of autonomous AI agents, enhancing workflows and decision-making. Combining technologies like Retrieval Augmented Generation (RAG), Knowledge Graphs, and foundation models offers scalable, cost-effective solutions for complex business cases.
4. **Real-Time and Creative Translation**:
   AI-powered real-time translation and creative localization are set to redefine communication

and storytelling. Attendees learned about the importance of combining human oversight with AI for nuanced tasks, such as adapting content for cultural relevance or marketing.

5. **Compliance and Security Are Non-Negotiable**:
   Stricter regulations like GDPR are reshaping AI strategies, emphasizing the importance of secure, compliant, and localized infrastructures. Organizations were advised to consider hybrid deployment models and proactive governance to avoid reputational and financial risks.

6. **Surprising Revelations**:
   One "aha moment" was the demonstration of AI's ability to transcreate highly emotive and culturally nuanced content, such as transforming telenovela synopses into engaging English scripts. Another was the realization that no-human-in-the-loop workflows could efficiently handle high-volume content, unlocking new monetization opportunities.

These takeaways highlighted AI's versatility and underscored the need for thoughtful integration to maximize its potential across industries.

# Quotes from Speakers

**Dion Wiggins (Host and Facilitator):**

- *"AI has become a universal necessity, embedded in everything from Google Maps to ChatGPT—it's not a question of if you'll adopt it, but how and when."*
- *"We're past the experiment phase; AI is now delivering real business value by improving efficiency, customer satisfaction, and innovation."*
- *"The next big leap is using AI to democratize innovation, enabling even smaller businesses to compete on a global scale with modular solutions."*
- *"The integration of AI into agile methodologies has matured, transforming software development and accelerating project delivery."*
- *"The biggest success stories in 2025 will come from businesses that align AI with measurable ROI rather than treating it as an experimental tool."*
- *"Tools like predictive analytics, chatbots, and recommendation engines are reshaping customer experience, creating satisfaction and loyalty at scale."*
- *"AI is not just about automation; it's about augmenting human creativity and decision-making to unlock new opportunities."*

**Professor Philip Kuhn (Subject Matter Expert):**

- *"The trend toward smaller, more efficient models shows that we can achieve high-quality results without massive computational costs."*

- *"Modular AI frameworks are transforming the landscape by combining RAG, Knowledge Graphs, and foundation models to outperform traditional custom solutions."*
- *"As algorithms become more efficient, and hardware advances, AI is becoming affordable for businesses of all sizes."*
- *"The competitive pressure to reduce costs and increase accessibility is driving innovation in edge AI and compact models."*
- *"Explainable AI and federated learning are helping to bridge the gap between innovation and compliance, especially in highly regulated industries."*
- *"Multimodal capabilities—integrating text, audio, and visual data—are redefining how AI interprets and interacts with the world around it."*
- *"Businesses must adopt AI strategically, focusing on modular and scalable solutions that can adapt to evolving needs in real time."*

**Dr. Joseph Sweeney (Industry Analyst, IBRS):**

- "The march of technology is predictable; the challenge lies in determining how and when organizations will adopt it effectively."
- "Generative AI is at the inflection point where its cost is low enough to become consumerized, much like wireless data revolutionized mobile technology."
- "AI agents are exciting, but their value will only be realized when they deliver measurable business outcomes and productivity gains."
- "Automation is where AI truly shines—streamlining processes and eliminating unnecessary work are key to proving its ROI."
- "Organizations that successfully orchestrate AI as a service will transform workflows and gain a significant competitive edge."
- "The current reliance on manual AI prompting is akin to using a computer without a mouse; it's clunky and inefficient."
- "2025 will be the year of orchestration—bringing together diverse AI tools to create seamless, automated systems."

# General AI Predictions

## Prediction: AI Will Prove Its Value in Everyday Business

In 2025, AI is expected to firmly establish its value in everyday business operations, moving beyond experimental phases to deliver measurable returns. This year will see AI solutions integrated more deeply into core business strategies, focusing on enhancing customer experience and streamlining processes. Businesses will increasingly rely on predictive analytics, chatbots, and recommendation engines to boost efficiency and customer loyalty.

This shift also marks the emergence of AI as an essential tool for fostering innovation and enabling competitive advantages. Companies that successfully adopt AI will benefit from increased productivity and cost savings, especially in areas like software development and agile project management. The emphasis will now be on execution, with best practices becoming more standardized.

- **Highlights from Transcript:**
    - Transition from experimentation to best-practice execution.
    - Failure rates of AI projects remain high but are declining as success stories emerge.
    - AI tools like predictive analytics, chatbots, and recommendation engines improve customer satisfaction.
    - Significant productivity gains observed, e.g., developers using tools like GitHub Copilot coding 10x faster.
    - Integration into agile methodologies is maturing.
- **Key Action Points:**
    - Shift focus from "cool" AI experiments to ROI-driven projects.
    - Invest in training staff on best practices for AI implementation.
    - Prioritize customer experience tools powered by AI for direct business impact.

## Prediction: AI Gets Affordable – The Coming Drop in Transaction Costs

The year 2025 is poised to be transformative as AI technology becomes more accessible and affordable for businesses of all sizes. With advancements in efficient algorithms, energy-saving hardware, and the availability of compact models, the cost of deploying AI solutions is expected to decrease significantly. These developments are democratizing AI adoption, enabling even small businesses to harness its potential.

Another factor driving this affordability is the adoption of modular AI architectures, such as RAG (Retrieval Augmented Generation) combined with pre-trained foundation models. This approach reduces the need for extensive customization, making AI both more scalable and easier to integrate into

existing workflows. Companies will find AI solutions to be a cost-effective way to enhance operations and remain competitive.

- **Highlights from Transcript:**
    - Model sizes are stabilizing or shrinking (e.g., Llama models) while maintaining or improving quality.
    - Cost reduction driven by efficient algorithms, energy-saving hardware, and compact models.
    - RAG (Retrieval Augmented Generation) and pre-trained models on platforms like AWS lower entry barriers for businesses.
- **Key Action Points:**
    - Leverage open-source and compact models for cost-efficient AI deployment.
    - Evaluate and adopt tools that combine RAG with foundation models for scalable solutions.
    - Explore renewable energy-driven AI solutions to reduce operational costs.

## Prediction: Foundation Model Leaders Shift Focus to AI Applications

As foundational models mature, 2025 will see leading organizations like OpenAI and Anthropic shifting their focus toward developing and monetizing application-layer AI solutions. This transition is driven by the need to generate sustainable revenue streams and differentiate themselves in a competitive market. These companies will prioritize user-facing applications that deliver immediate business value.

The financial dynamics of AI are also changing. While the hardware layer has traditionally been the most profitable, the application layer is set to take the lead, mirroring the evolution of the mobile app ecosystem. This shift will spur innovation in AI applications across industries, creating new opportunities for businesses to leverage cutting-edge technologies.

- **Highlights from Transcript:**
    - Frontier labs like OpenAI face high training costs, leading to a shift toward application-based revenue streams.
    - Applications such as cloud-based AI services are mirroring trends seen in the mobile app ecosystem.
- **Key Action Points:**
    - Monitor emerging application-specific AI products.
    - Position your business to integrate AI-driven applications that align with your operations.

- Strategize partnerships with AI vendors to gain early access to application ecosystems.

## Prediction: 2025 Will Be he Year Of AI Agents

AI agents are expected to revolutionize industries in 2025 by automating workflows, improving decision-making, and delivering personalized experiences at scale. These agents, capable of reasoning and self-correction, will enable businesses to streamline complex processes and reduce manual intervention in repetitive tasks.

The transition from standalone tools to orchestrated ecosystems of agents will redefine operational efficiencies. From customer support to internal task automation, AI agents will become indispensable in driving business transformation. Their ability to operate autonomously while integrating seamlessly with existing systems will be a game-changer.

- **Highlights from Transcript:**
  - Agents automate workflows, enhance decision-making, and deliver personalized experiences.
  - Agents' capabilities include reasoning, execution, and self-correction.
  - Real-world use cases include task automation and customer support.
- **Key Action Points:**
  - Begin integrating AI agents in customer service and internal processes.
  - Train employees on orchestrating AI agents to maximize efficiency.
  - Experiment with task-specific agents to automate repetitive workflows.

## Prediction: RAG + Knowledge Graphs + Agents + Foundation Models Will Outperform Custom Models

In 2025, businesses will increasingly adopt modular AI frameworks combining RAG, Knowledge Graphs, and foundation models to solve complex challenges. This modular approach will outperform traditional custom models by offering flexibility, scalability, and cost-efficiency. These solutions allow real-time updates and provide structured reasoning capabilities, making them highly adaptable to changing business needs.

The modular architecture also enhances deployment speed and reduces resource consumption, making AI accessible to a wider audience. Businesses adopting this approach will benefit from reduced operational costs and faster time-to-market for AI-driven solutions.

- **Highlights from Transcript:**
  - Modular AI outperforms resource-intensive custom models in cost-efficiency and scalability.
  - Dynamic updates using RAG and Knowledge Graphs ensure relevance and accuracy.

- **Key Action Points:**
    - Build AI solutions using modular frameworks.
    - Use RAG and Knowledge Graphs for real-time insights and structured reasoning.
    - Focus on API-based pre-trained models for faster deployment.

## Prediction: Regulations, Security, Sovereign Data, and Compliance Will Drive AI Strategies

In 2025, the regulatory landscape will play a crucial role in shaping AI strategies across industries. Stricter laws such as GDPR and data sovereignty regulations will push businesses toward more secure and compliant AI implementations. This trend highlights the importance of proactive governance and transparency in building consumer trust.

Organizations will prioritize secure deployment models, including on-premise and hybrid cloud solutions, to address compliance challenges. Advanced cybersecurity measures and localized infrastructures will also be key to navigating this evolving landscape while ensuring long-term success.

- **Highlights from Transcript:**
    - Stricter regulations like GDPR and data sovereignty laws are reshaping AI deployment.
    - Penalties for non-compliance are substantial, emphasizing the need for proactive governance.
    - On-premise and hybrid cloud models are gaining popularity for secure AI applications.
- **Key Action Points:**
    - Conduct compliance audits to align with regional regulations.
    - Invest in secure on-premise or hybrid solutions to ensure data control.
    - Build transparency mechanisms into AI workflows to enhance consumer trust.

## Prediction: From Cloud-First to Control-First – The Evolution of AI Infrastructure

In 2025, a significant shift from cloud-first to control-first strategies will redefine AI infrastructure. Businesses will increasingly adopt self-hosted and private cloud solutions, driven by the need for greater data control, customization, and compliance with stringent regulations. These models offer enhanced security while maintaining scalability and operational efficiency.

This evolution will pave the way for hybrid solutions that balance the advantages of on-premise systems with the flexibility of cloud-based services. Companies that prioritize control-first strategies will be better positioned to address emerging compliance requirements and leverage AI effectively.

- **Highlights from Transcript:**
    - Surge in self-hosted and private cloud AI solutions for better control and customization.

  - Strategic advantages include compliance with data sovereignty laws and operational scalability.

- **Key Action Points:**
  - Transition critical AI workloads to private cloud or on-premise setups where compliance is a priority.
  - Evaluate hybrid solutions for scalable AI integration.
  - Educate IT teams on managing localized infrastructures and federated learning systems.

# Language Processing, Localization, and Translation Predictions

## Prediction: AI Frequently Outperforms Humans in Language-Related Work and Processes

In 2025, AI will continue to surpass human capabilities in specific language-related tasks, particularly those involving large-scale, repetitive, or time-sensitive processes. AI-driven solutions are expected to excel in areas like content generation, bulk translations, and automated quality assurance, delivering unmatched speed and scalability.

While humans remain crucial for nuanced and creative tasks, AI's ability to handle domain-specific challenges such as legal and medical content will further establish its dominance. This trend will drive a collaborative approach where AI handles routine work, and humans focus on higher-value contributions.

- **Highlights from Transcript:**
  - AI excels in bulk content handling, real-time updates, and domain-specific tasks (e.g., law, medicine).
  - Automated QA offers significant advantages over human spot checks.
- **Key Action Points:**
  - Use AI for repetitive tasks such as bulk translation and transcription.
  - Integrate automated QA tools to enhance quality and reduce errors.
  - Focus human resources on nuanced and creative tasks.

## Prediction: AI-Assisted Translation Will Be the Norm

In 2025, AI-assisted translation will become the industry standard, streamlining workflows and enhancing localization efforts. Advanced tools incorporating automated post-editing and quality assurance capabilities will enable faster project completion and higher accuracy.

This shift will redefine the localization industry, with AI supporting every stage of the translation process, from research and terminology extraction to delivery. Businesses leveraging these tools will achieve better scalability and cost-efficiency while maintaining high-quality outputs.

- **Highlights from Transcript:**
    - Enhanced workflows with automated post-editing and QA steps.
    - Seamless integration of AI into every stage of localization, from preparation to delivery.
- **Key Action Points:**
    - Adopt AI-assisted translation tools with built-in post-editing capabilities.
    - Optimize workflows to balance machine efficiency with human creativity.
    - Train teams on leveraging modular AI systems for large-scale localization projects.

## Prediction: AI-Powered Conversations and Real-Time Translation Will Be Everywhere

In 2025, AI-powered real-time translation will become ubiquitous across devices and platforms, enabling seamless communication across languages. Integration with conversational AI will provide users with natural and intuitive interactions, transforming industries like customer service, healthcare, and education.

These advancements will redefine multilingual accessibility and collaboration, making real-time translation a cornerstone of global communication. Businesses that embrace this technology will gain a competitive edge in catering to diverse audiences and markets.

- **Highlights from Transcript:**
    - AI-driven real-time translation across devices (smartphones, glasses, etc.).
    - Integration with conversational AI for seamless user experiences.
- **Key Action Points:**
    - Develop or adopt solutions for real-time multilingual support.
    - Train support teams to utilize AI-powered conversational interfaces.
    - Explore opportunities in industries like healthcare, government, and customer service.

## Prediction: Widespread Use of AI in Creative Translation and Localization

Creative translation and localization will be revolutionized in 2025 as AI takes on more artistic and emotionally resonant tasks. From adapting marketing campaigns to writing synopses, AI's ability to capture tone, intent, and cultural nuances will become a key asset for businesses.

Collaboration between humans and AI will be essential to ensure that creative outputs maintain authenticity and align with specific cultural contexts. This synergy will unlock new possibilities for storytelling and audience engagement in diverse markets.

- **Highlights from Transcript:**

  - AI aids in transcreation by adapting tone, intent, and cultural nuances.
  - AI tools assist in storytelling and creating emotionally resonant content.
- **Key Action Points:**
  - Leverage AI for marketing campaigns that require cultural adaptation.
  - Use AI to optimize creative workflows, including content generation and style tuning.
  - Collaborate with AI to enhance human creativity in localization tasks.

### Prediction: The Rapid Rise of No-Human-In-The-Loop Translation

The rise of no-human-in-the-loop translation in 2025 will transform how businesses handle high-volume translation projects. Automated workflows, enhanced by QA and post-editing algorithms, will enable faster turnaround times and significant cost savings without sacrificing quality.

While this approach may not suit every use case, it will be a game-changer for applications such as media localization, real-time subtitling, and large-scale content processing. Organizations adopting this technology will unlock new monetization opportunities and reach broader audiences.

- **Highlights from Transcript:**
  - End-to-end workflows with automated QA reduce human involvement.
  - Applications include bulk media processing and real-time subtitling.
- **Key Action Points:**
  - Deploy no-human-in-the-loop workflows for high-volume projects.
  - Focus on use cases where speed and cost-efficiency outweigh perfection.
  - Monitor and address areas where human intervention might still be necessary.

# Questions and Answers

The Q&A session explored critical questions about the advancements, challenges, and applications of AI in text recognition, translation, and accessibility. Topics ranged from leveraging AI for low-resource languages and integrating translation memories into workflows to the evolving role of CAT tools and real-time translation's impact on accessibility. These insightful discussions underscored AI's transformative potential and provided practical guidance for businesses looking to adopt and optimize AI technologies.

**Q1: How is text recognition coming along? We have a large set of data confined to PDFs, and it would do wonders if these could be accessed for predictive purposes.**
**A1:** Text recognition has seen significant advancements, particularly with the integration of AI models that combine text and vision capabilities. These models can effectively process PDFs, including

extracting text from challenging formats like tables or scientific documents. Tools like ChatGPT and OCR-enhanced AI systems can handle character recognition while understanding broader context, making it easier to extract meaningful data from PDFs.

**Q2: How many languages does AI voice response currently include, and are there plans to support all languages? Can it handle slang effectively?**
**A2:** AI voice response currently supports 55 languages for interaction and can synthesize speech in around 140 languages. It is trained to recognize and handle slang, though glossaries and customization may be required for specific jargon or regional colloquialisms. For example, specialized glossaries can be created to accurately process names, terminology, or regional idioms in real-time applications.

**Q3: Should we still focus on learning and training with standard CAT tools, given the rise of AI-assisted translation?**
**A3:** Yes, CAT tools are not becoming obsolete. Instead, they are evolving to become smarter and better integrated with AI. These tools will continue to play a crucial role in managing workflows, terminology, and quality assurance, providing a framework where AI can be effectively utilized alongside human oversight.

**Q4: Are there existing platforms for multi-modal translation, and how developed are they?**
**A4:** Multi-modal translation platforms like ChatGPT are already operational, integrating speech, text, and visual inputs. Other efforts, such as Google's AudioPalm and Meta's Seamless MT, are in development, aiming to create end-to-end solutions. While many of these systems are still maturing, they are making significant progress in enabling seamless communication across modalities.

**Q5: With accessibility regulations increasing, how can real-time translation address these requirements?**
**A5:** Real-time translation can support accessibility by generating captions, transcriptions, and audio outputs for people with hearing or visual impairments. AI systems are being developed to recognize not only spoken words but also tonal nuances, emotional cues, and background sounds, ensuring compliance with accessibility standards and improving user experience.

**Q6: Current large language models perform well in high-resource languages but struggle with low-resource ones. What are the solutions or predictions for this issue?**
**A6:** Addressing low-resource language challenges involves government and private sector initiatives to fund large language model development tailored to these languages. Techniques like leveraging monolingual data, synthetic data generation, and balancing training datasets are being explored. Future advancements aim to deliver consistent performance across all languages by ensuring better cross-linguistic transfer of knowledge.

**Q7: Can translation memories be used as part of retrieval-augmented generation (RAG) for AI systems?**
**A7:** Absolutely. Translation memories can enhance AI systems by serving as a structured database for terminology, usage examples, and domain-specific context. This approach is already in use, with AI leveraging translation memories to provide accurate and contextually relevant outputs, especially in specialized fields like legal or medical translation.

**Q8: Will CAT tools become redundant with the rise of AI?**
**A8:** No, CAT tools will not become redundant. They will evolve to incorporate AI capabilities, enhancing features like interactive translations, glossary management, and quality control. The synergy between CAT tools and AI will provide better user interfaces and improve translation workflows.

**Q9: What improvements can be expected in handling low-quality PDFs with AI?**
**A9:** AI advancements, particularly in OCR and vision-based language models, are enabling better handling of low-quality PDFs. These systems can synthesize missing or unclear text and interpret context to fill gaps, making data extraction from poorly formatted or scanned documents much more accurate and reliable.

**Q10: What steps should businesses take to integrate real-time translation into their workflows?**
**A10:** Businesses should begin by identifying key use cases where real-time translation can add value, such as customer support or accessibility compliance. Investing in multi-modal platforms and training staff to leverage conversational AI interfaces will ensure smoother adoption. Customizing solutions with glossaries and domain-specific adaptations can further enhance the effectiveness of real-time translation systems.

# Key Takeaways

The webinar on **AI Predictions for 2025** provided attendees with a comprehensive overview of the transformative trends shaping the AI landscape. Below are the most significant insights and actionable points derived from the session:

**1. AI's Transition from Experimentation to Essential Business Tool**

- **Insight**: AI has moved past its experimental phase to become a cornerstone of business strategy, delivering measurable ROI and improving efficiency across industries. Tools like predictive analytics, chatbots, and recommendation engines are enhancing customer experience and driving innovation.
- **Actionable Advice**: Businesses should focus on implementing AI projects that are ROI-driven and align with clear operational goals.

**2. The Democratization of AI**

- **Insight**: Cost reductions driven by efficient algorithms, compact models, and energy-saving hardware are making AI accessible to businesses of all sizes. Open-source models and platforms like AWS Bedrock are enabling even small enterprises to leverage AI effectively.
- **Actionable Advice**: Explore affordable, modular AI solutions to scale operations while minimizing upfront investment.

**3. Modular AI as the Future**

- **Insight**: Modular AI frameworks, combining RAG, Knowledge Graphs, and foundation models, are more efficient, scalable, and cost-effective than custom-built models. These systems enable real-time updates and structured reasoning.
- **Actionable Advice**: Prioritize modular architectures to reduce resource consumption, accelerate deployment, and maintain flexibility.

**4. The Rise of AI Agents**

- **Insight**: 2025 will see widespread adoption of AI agents capable of reasoning, executing tasks autonomously, and delivering hyper-personalized experiences. These agents will redefine workflows and enhance decision-making.
- **Actionable Advice**: Begin experimenting with task-specific AI agents to automate workflows and improve operational efficiency.

**5. Real-Time Translation and Conversational AI Everywhere**

- **Insight**: AI-powered real-time translation and conversational interfaces are becoming ubiquitous, transforming industries like customer service, healthcare, and government. These technologies enhance multilingual communication and accessibility.
- **Actionable Advice**: Integrate real-time translation tools and conversational AI into workflows to enhance global reach and user experience.

**6. The Balance Between AI and Human Creativity**

- **Insight**: AI is excelling in creative tasks such as transcreation, where it adapts tone, intent, and cultural nuances. However, collaboration with humans remains essential to ensure authenticity and artistic relevance.
- **Actionable Advice**: Leverage AI for content generation and creative localization while maintaining human oversight for nuanced and cultural-sensitive tasks.

**7. Compliance and Security Are Foundational**

- **Insight**: With stricter regulations like GDPR and increased focus on data sovereignty, businesses must adopt secure deployment models such as on-premise and hybrid solutions. Non-compliance risks significant financial and reputational damage.
- **Actionable Advice**: Conduct compliance audits, adopt secure infrastructures, and implement proactive governance to ensure AI strategies align with regulatory demands.

**8. The Rise of No-Human-In-The-Loop Workflows**

- **Insight**: End-to-end AI workflows are enabling high-quality, large-scale translation and localization projects without human intervention. While not suitable for all scenarios, they are unlocking new monetization opportunities.

- **Actionable Advice**: Deploy no-human-in-the-loop solutions for high-volume tasks and real-time media localization to reduce costs and increase scalability.

**Surprising Revelations and "Aha Moments"**

- **AI Creativity Surpassing Expectations**: Examples of AI creatively adapting telenovela synopses for English audiences demonstrated how machines can add emotional depth and dramatic flair.
- **AI-Powered Research Chains**: One speaker shared how AI orchestrations reduced a two-week research process to just an hour, exemplifying its potential for streamlining complex tasks.
- **Real-Time Interaction Potential**: The live demonstration of conversational AI hinted at the transformative possibilities of real-time, voice-enabled interactions across devices.

# Speaker Analysis

## Dion Wiggins

**Role**: Host and Facilitator
**Profile**:
Dion serves as the central facilitator of the presentation, guiding the audience through the predictions and ensuring a smooth flow of discussions. As a knowledgeable and engaging speaker, Dion emphasizes the practical applications of AI across industries. He combines technical insights with business foresight, sharing examples from his personal experiences, such as using GitHub Copilot and orchestrating AI tools in his team. Dion's approach is collaborative, often inviting other speakers to share their expertise while ensuring audience engagement through Q&A and practical examples.

**Key Contributions**:

- Framed the predictions within the context of business and operational impacts.
- Shared practical insights into AI integration, such as productivity gains from tools like Copilot.
- Highlighted advancements in customer experience, security, and compliance as critical drivers for AI adoption.
- Demonstrated a clear understanding of strategic industry needs, offering actionable steps for businesses to align with AI trends.

## Professor Philip Koehn

**Role**: Subject Matter Expert on Technology and AI Trends
**Profile**:
Based in Bangkok, Professor Philip Koehn brings a deeply technical perspective to the presentation,

focusing on trends in model efficiency, cost reduction, and the modularity of AI systems. With a background in research and a collaborative spirit, Philip delves into the mechanics of AI development, such as model sizes, algorithm efficiency, and the competitive landscape. His ability to articulate complex technical topics for a broad audience is one of his strengths, supported by examples from industry trends and emerging technologies.

**Key Contributions**:

- Provided detailed insights into AI model optimization, such as quantization and energy-saving hardware.
- Highlighted the growing importance of modular AI architectures like RAG and Knowledge Graphs.
- Emphasized the potential of compact, efficient models for democratizing AI access.
- Brought a global perspective on AI adoption, reflecting regulatory and competitive pressures.

## Dr. Joseph Sweeney

**Role**: Research Director and Advisor on the Future of Work, Guest Analyst Speaker
**Profile**:
Doctor Joseph Sweeney offers a high-level, strategic perspective, blending decades of experience with a focus on how emerging AI technologies impact business operations. Speaking via video, Joseph brings a historical lens to the evolution of AI and emphasizes the need for orchestration and automation over manual prompting. His pragmatic view highlights challenges in AI adoption while emphasizing the importance of long-term planning and innovation in workflow automation.

**Key Contributions**:

- Advocated for a shift from manual AI prompting to seamless orchestration in business processes.
- Explained the financial and operational challenges facing AI vendors and their enterprise clients.
- Shared thought-provoking examples of AI-driven research processes that streamline operations.
- Predicted the rise and potential pitfalls of AI agents and their applications in real-world scenarios.

## Humphrey Bot

**Role**: Demonstration AI Agent
**Profile**:
Humphrey Bot serves as a live demonstration of conversational AI capabilities, showcasing real-time information retrieval and task execution. Designed as an interactive tool, Humphrey Bot exemplifies

how AI agents can assist in business operations, such as conducting research, answering customer queries, and performing data-driven tasks. Although technical issues affected the live demo, Humphrey Bot illustrates the potential of AI-driven conversational interfaces to revolutionize user interactions and enhance productivity.

**Key Contributions**:

- Demonstrated the capabilities of AI agents to conduct real-time research and deliver concise outputs.
- Highlighted the integration of conversational AI into workflows, showcasing its relevance for customer support and operational tasks.
- Served as an example of seamless AI-human collaboration in task management, despite technical limitations during the session.

# Full Presentation Transcription

This transcript is a verbatim record of the spoken words, captured exactly as they were said. It has not been edited to correct grammar or remove disfluencies.

| Time | Speaker | Transcription |
|---|---|---|
| 00:00:07 | Dion | Okay. Hello everyone, and welcome to the omniscient presentation on AI and Language Processing Predictions for 2025. |
| 00:00:15 | Dion | I'm joined today by Professor Philip Koehn, live in Bangkok. He's sitting beside me and we'll also be joined by Doctor Joseph Sweeney. |
| 00:00:24 | Dion | Unfortunately, it's 2 a.m. where Joe is, so he is gratefully prepared a video for us that we will play that will talk a little bit about one of his key focus areas. |
| 00:00:38 | Dion | It's important to note that there is an extensive series of blog posts that will be published on Monday, over 100 pages of content that is backing all of this up with research. Of course, there's the summaries as well. |
| 00:00:51 | Dion | And today we'll just be going through all the summaries on these key predictions. |
| 00:00:55 | Dion | If you have any questions, please don't hesitate to ask using the chat, and we will be presenting and answering those questions mostly at the end. |
| 00:01:07 | Dion | Also, this video will be available by recording. So, if you have to drop off because it is a 90-minute video today, if you have to drop off, that's fine. |

| | | |
|---|---|---|
| 00:01:16 | Dion | You'll be able to pick it up later on. You'll also be able to share it with your friends as needed. |
| 00:01:21 | Dion | Okay, so we're going to get things started. |
| 00:01:24 | Dion | And let's start with our first a set of general AI predictions. And then second we will follow with language industry and language processing specific AI predictions. |
| 00:01:36 | Dion | Okay. So, the first prediction is that AI will prove its business value in everyday business. |
| 00:01:44 | Dion | Now we're past a phase where AI, for AI's sake has been happening. It's been constant. |
| 00:01:51 | Dion | We've had a lot of our AI just adopted by people, and it hasn't really gone well in many cases the business has come down and said, hey, let's go and do I. This sounds cool. |
| 00:02:02 | Dion | All these promises, all these things, they've done it poorly. |
| 00:02:06 | Dion | And in in the article blog post, you'll see more details on why projects are extremely high. Failure rates for I more than double traditional failure rates, but what we're seeing now is we're moving away from experimentation to real, solid, best practice driven execution. |
| 00:02:26 | Dion | Okay, you're going to see a lot more success stories. You're going to see a lot more ROI studies come out. |
| 00:02:33 | Dion | The early challenges have basically been passed. We're starting to see real business value. |
| 00:02:38 | Dion | And we're addressing also persistent challenges overall. |
| 00:02:44 | Dion | So, you know, there's a lot going on and you're going to see these changes very quickly. |
| 00:02:51 | Dion | A lot of it is about customer experience. |
| 00:02:54 | Dion | And with that customer experience, of course, that comes down to everything from predictive analytics, chatbots, recommendation engines, and overall increasing satisfaction and loyalty. |
| 00:03:08 | Dion | The industry specific applications is a good list here. I won't read them all, but you get the general idea. |
| 00:03:14 | Dion | These slides, of course, will be available later because they are very rich in text along with all the other information. Okay, so one of the things that's really notable is the revolution in software development. |
| 00:03:32 | Dion | So, from things like Copilot's that are connecting to GitHub. |
| 00:03:35 | Dion | I mean, my team here in Bangkok. We use GitHub and we use copilot heavily. |

| | | |
|---|---|---|
| 00:03:41 | Dion | Just about every piece of new code we're writing has extensive AI backing to it. And I'm still coding, even though I'm the CTO and I'm coding ten times faster, maybe even more than I've ever coded before. |
| 00:03:56 | Dion | So, it's quite substantial. |
| 00:03:58 | Dion | Also, being able to just say, I know how to do it, but show me and you know. Show me. |
| 00:04:03 | Dion | Is this good? Check my code, add error handling, all of these kinds of things. |
| 00:04:07 | Dion | How is somebody else done this complex task? Or analyze my code and find the bugs? |
| 00:04:12 | Dion | These kinds of things are all very straightforward. Now with AI, we're also getting to the point where it's integration with existing methodologies has started to mature. |
| 00:04:23 | Dion | So, agile has really adopted, well, AI. And we're moving into agile based AI development. |
| 00:04:32 | Dion | There's a lot of areas where we're going to see key advancements we already have. So, things like multi modal which Philip will be talking a little bit on later. |
| 00:04:41 | Dion | Things like human like reasoning natural interfaces. And we'll give you some examples of natural interfaces shortly. |
| 00:04:49 | Dion | Federated learning overall, just generative AI. As you've seen, Philip will mention a little bit later on also about explainable AI. |
| 00:04:58 | Dion | And of course, don't forget a very important part of the equation is edge AI, where the processing is actually moving out to the edge. And there's even chips now that are specialized for running AI tasks on your mobile devices, internet of things, cell phones, etc.. |
| 00:05:15 | Dion | Okay, so the bottom line is that AI is becoming an essential business tool in driving efficiency, growth, and creating opportunities across industries. We're past the experiment phase, so expect to see some of the failure rates drop, but it's still going to be very high as many companies still don't yet have the experience. |
| 00:05:35 | Dion | But that will go hand in hand with success stories. |
| 00:05:41 | Dion | Okay, so I'm going to pass over to Philip now to talk about affordable AI. |
| 00:05:45 | Philipp | Yeah. So, I will mainly talk about like the trends both in terms of technology and business that cause AI to become more affordable. |
| 00:05:52 | Philipp | As we all know, we're dealing here with gigantic models that have really become too big. |

| | | |
|---|---|---|
| 00:06:00 | Philipp | They are too big to train. They cost tens of millions of dollars to train. |
| 00:06:03 | Philipp | They're too big to use as really high inference cost and fairly expensive GPU setups. You need to run them. |
| 00:06:09 | Philipp | And they also then too big to adapt to use cases. So, this is actually a chart from like almost two years ago. |
| 00:06:15 | Philipp | Now what the size of the models. And you see the kind of max out at around 500 billion parameters. |
| 00:06:20 | Philipp | So, the good news already is the models didn't actually get bigger than that. |
| 00:06:24 | Philipp | So, now we also still at most talk about, you know, for instance, the largest llama model is about 400 billion parameters. And you rather have a trend towards smaller models. |
| 00:06:33 | Philipp | So, this is already happening. |
| 00:06:38 | Philipp | Okay. Here are just a few more examples about like cost, training cost and the compute involved in training these models to just very specific examples that We found some numbers for the. |
| 00:06:48 | Philipp | Google's Gemini costs $191 million to train GPT four. 78 million to train, according to some models. |
| 00:06:55 | Philipp | Just to kind of give you a comparison. The original transformer model that was originally proposed for machine translation, if you look at that architecture and you would train it on the on the data we have, it would only cost about 900 $900. |
| 00:07:09 | Philipp | Okay. |
| 00:07:10 | Philipp | One interesting thing that is happening in this space is that we actually have different clusters emerging in this trade off of quality and cost. So, here you see a chart where on the x axis you see the price average price per token going from 1 to 50. |
| 00:07:29 | Philipp | And the other one is like some measure of quality which is the chatbot arena. |
| 00:07:35 | Philipp | And so there is now already the tendency that multiple models of multiple sizes are being offered at different price points. |
| 00:07:45 | Philipp | So, what will also drive making AI more affordable is various key drivers. One is more efficient algorithms. |
| 00:07:54 | Philipp | There's clearly a lot of interest in coming up with more efficient algorithm. Things like quantization and Elmo and so on. |
| 00:08:02 | Philipp | There's also more energy saving hardware. So, we see now also GPUs coming available from AMD. |

| | | |
|---|---|---|
| 00:08:10 | Philipp | And clearly we hope that maybe over the medium term time horizon there will be more competition in that space. And that will drive costs for GPUs down substantially. |
| 00:08:20 | Philipp | There are a lot of compact and efficient models available, even for free. I mean, there's clearly the big efforts for meta to open source release models, and there's also just kind of the competitive industry pressure of putting AI into edge devices and therefore really, really a lot of effort and push towards cost reduction. |
| 00:08:41 | Philipp | So, we expect Dramatic cost reductions over time. |
| 00:08:46 | Philipp | And also then that kind of has benefits for kind of all sizes of models you want to use. And all kinds of companies want to use this model for various purposes. |
| 00:08:56 | Dion | And I'll just chip in here to one of the things we've been using a lot of the meta models as well. And, you know, not long ago there was llama 3.14 or 5 B. |
| 00:09:08 | Dion | Now that's a huge model. You need huge hardware and it's very expensive to run comparably. |
| 00:09:14 | Dion | Meta's llama 3.37 TB much smaller model, and it's already delivering higher quality results than the much larger model was delivering. |
| 00:09:26 | Philipp | So, there's actually a lot of competitive pressure in this space to drive down costs. So, we actually have several companies that build large language models. |
| 00:09:33 | Philipp | I mean, there's clearly almost like every month there's a new large language model comes around. |
| 00:09:38 | Philipp | And what pressures Of these. A possibility of these companies to charge a lot for AI access is also freely available. |
| 00:09:46 | Philipp | Models like llama, and there are many ongoing efforts from the government funding and nonprofits to build models that are then freely available. |
| 00:09:56 | Philipp | Nobody really has a secret sauce. All these models look very, very similar and build very, very similar, built on pretty much all available data that people can get their hands on. |
| 00:10:06 | Philipp | And so there's no really huge competitive advantage to any of them at the time. |
| 00:10:14 | Philipp | On the other hand, there's also clearly the drive from the user cost to minimize expenses. There is always a trade off between quality and price. |

| | | |
|---|---|---|
| 00:10:23 | Philipp | And there's also several ways to balance maybe diminished quality at a cheaper price point with a lot of task specific adjustments. We're going to talk much more about that later. |
| 00:10:37 | Philipp | Okay. |
| 00:10:43 | Philipp | One. Okay. |
| 00:10:44 | Dion | Okay, I'll take this one. Philip. |
| 00:10:46 | Dion | So, one of the things that definitely decreasing costs is the use of tools such as rag retrieval, augmented generation, and knowledge graphs in combination with language models. Now, one of the things that's getting really noticeable is the quality that even the the free models, such as llama 3.3 is producing is adequate for a huge range of tasks. |
| 00:11:14 | Dion | So, combining Rag and Knowledge Graph often can mean you don't need to spend time doing fine tuning anymore. |
| 00:11:21 | Dion | And it has many other benefits as well, such as keeping data up to date because you can just update Rag data very quickly rather than having to constantly fine tune. |
| 00:11:31 | Dion | There's also some great new models on hugging face, and they're constantly getting new models, all sorts of things that, you know, are really targeted in different areas. |
| 00:11:40 | Dion | One of the interesting ones is big Science is Bloom. That's really focused on multilingual support across the board with a huge range of languages. |
| 00:11:50 | Dion | And then the last point on this slide is all about easy to access. So, you've got pre-trained models of course, that we've just talked about. |
| 00:11:57 | Dion | But services from companies like AWS with their bedrock service where you can just call language models on demand and pay for a few transactions. |
| 00:12:05 | Dion | And they're a fraction of the cost of something like ChatGPT. |
| 00:12:09 | Dion | They really are low cost. We use them heavily. |
| 00:12:13 | Dion | And then you've got, you know, various different ways to do strategic resource allocation to make sure those things go smooth and using renewable energy. Everything's starting to come down across the board. |
| 00:12:24 | Dion | I'll be talking a little bit more later about where things are moving with that and why they're coming down across the board as well. |
| 00:12:33 | Dion | So, let's take this to its conclusion. |
| 00:12:37 | Dion | One of the things is, is really and truly democratization of innovation. |

| | | |
|---|---|---|
| 00:12:42 | Dion | You know, these are Lego blocks. They plug in, smaller businesses can join in, they can try things, they can innovate without having to have all the expense of these kinds of models being built, or even the complexity. |
| 00:12:55 | Dion | Just call an API in many cases, of course, that has a much broader impact across every industry in the world. |
| 00:13:03 | Dion | And let's be honest, AI is everywhere now. It's ubiquitous. |
| 00:13:06 | Dion | It's a universal necessity. The things we do, whether we're using Google Maps, it's using AI in the background, whether we're using ChatGPT or whether we're using a range of other things. |
| 00:13:17 | Dion | All these tools have become ubiquitous. So, at this point, it's not a matter of if you want to adopt, it's a matter of how or when and then how are you going to leverage it. |
| 00:13:31 | Dion | Okay. |
| 00:13:32 | Dion | So, moving on to a really interesting one. So, foundation model leaders are going to shift their focus to AI applications for growth and sustainability. |
| 00:13:42 | Dion | Now that sounds a little odd, but I'll justify this in a couple of slides time and you'll see why. |
| 00:13:47 | Dion | So, the frontier labs like OpenAI and anthropic, they are spending enormous amounts of money on training foundational models. It's extremely competitive and they have to stay on top of it. |
| 00:14:00 | Dion | You can't just sleep even for a day because the next model comes out. |
| 00:14:04 | Dion | So, there needs to be new revenue streams and new sustainable business models overall. |
| 00:14:11 | Dion | You know, we need to you know, the competition there is massive and it's intensifying and there's diminishing returns. You know the models are getting so good in so many areas. |
| 00:14:22 | Dion | Now let's look at the other side of it. Where is the money really being made today. |
| 00:14:27 | Dion | Well about 83% of the money in the AI space is at the bottom at the chips and the semiconductors and GPUs. |
| 00:14:35 | Dion | Right? It's not actually in the actual models themselves. |
| 00:14:41 | Dion | So, there's a lot of challenges for frontier model development. As I said, diminishing returns. |
| 00:14:46 | Dion | The costs are excessive. |

| | | |
|---|---|---|
| 00:14:49 | Dion | You know, some of these Nvidia GPUs are anywhere from 30 to $80,000. |
| 00:14:54 | Dion | And you need thousands of them. Like I just read a few days ago that GPT four is trained on 100,000 GPUs. |
| 00:15:04 | Dion | That's something like $3 billion worth of hardware. It's significant. |
| 00:15:09 | Dion | So, there's some really good reasons why these companies are going to start. And they already have, in many ways, moving towards applications on top of their models. |
| 00:15:19 | Dion | Now let's take a look at where the money is being made. |
| 00:15:22 | Dion | So, you can see in the app space $400 billion. Semiconductors. |
| 00:15:27 | Dion | It's down the bottom right. But in generative AI, it's inverted. |
| 00:15:31 | Dion | $5 billion is being made at the top right. And at the bottom, it's all about Nvidia, right? |
| 00:15:36 | Dion | So, there's a lot of money being made in the hardware space right now. |
| 00:15:40 | Dion | But that will change. |
| 00:15:42 | Dion | Okay. Now if you break that down to the next frontier, this is how it's going to start moving. |
| 00:15:48 | Dion | So, you can see those changes. |
| 00:15:50 | Dion | You can see where the hardware is differing and where the applications are percentage wise are making money. |
| 00:15:57 | Dion | So, this is what you should expect to see. |
| 00:16:00 | Dion | You know look at cloud over the last ten years. You know it started with everything driving. |
| 00:16:05 | Dion | The cloud was hardware driven right. Then the infrastructure on top like AWS Azure, right. |
| 00:16:11 | Dion | And today, you know, all of those spaces have grown. But the real money is at the top. |
| 00:16:16 | Dion | And the apps, that's where you make the biggest money. And that's the same thing that's going to happen with generative AI. |
| 00:16:22 | Dion | So, the hardware is going to become less of a primary component. And the applications across the board, no matter what they may be. |
| 00:16:29 | Dion | So, you know, there's higher profit margins. There's greater differentiation, you know, so basically they're going to move up the stack. |

| | | |
|---|---|---|
| 00:16:38 | Dion | And this mirrors the trends of mobile phones for example. It used to be all the money was in the phones. |
| 00:16:43 | Dion | Now look at all the app stores. Apple makes a lot more money, you know, simply out of their app store by charging commissions. |
| 00:16:49 | Dion | Okay. |
| 00:16:50 | Dion | So, you know there's going to be some challenges. There's going to be a lot of competition. |
| 00:16:54 | Dion | And they may even alienate some customers. And there's various kinds of complexities. |
| 00:16:58 | Dion | But it is going to happen. |
| 00:17:01 | Dion | Okay. Now this is one of the really interesting ones agents. |
| 00:17:05 | Dion | So, agents will redefine industries by automating workflows, enhancing decision making and delivering hyper personalized experience across the board. Now, I just had an ad on Instagram two days ago. |
| 00:17:20 | Dion | I was in the ad. |
| 00:17:21 | Dion | It modified me to be in a video right now. that's pretty scary and pretty interesting and impressive all at the same time. |
| 00:17:30 | Dion | So, we're also going to transition from standard tools to overall orchestrated ecosystems of agents. So, let's explore this. |
| 00:17:38 | Dion | What is an agent and what can it do. |
| 00:17:40 | Dion | So, an agent can complete tasks all by itself, creating its own prompts and getting the job done. It can generate complete tasks on a goal or based on a goal. |
| 00:17:50 | Dion | And I put a goal down the bottom, for example. |
| 00:17:54 | Dion | It can then figure out if it's done a good job. If it can complete the task, it can read, write, and execute code. |
| 00:18:01 | Dion | In fact, I used some technology today to actually generate some of these slides, including the PowerPoint itself, including the text in it, and then just tweaked it a little and updated it. |
| 00:18:13 | Dion | So, what this basically means though, is it can think for itself. It can reason with itself, act on its own, and then decide if it made a good or a bad decision. |
| 00:18:22 | Dion | So, here's a really simple goal that you could give it. Analyze the latest news for data privacy issue relating to public AI services such as ChatGPT, and publish it on the omniscient WordPress website. |
| 00:18:34 | Dion | That's it. It will do every single step for you. |

| | | |
|---|---|---|
| 00:18:37 | Dion | It'll work out how to do it and move along. |
| 00:18:39 | Joe | Hi, my name is Doctor Joe Sweeney. I am the research. |
| 00:18:42 | Dion | Okay. One moment. |
| 00:18:44 | Dion | We just have to go back and get Joe's audio. He seems to be a little silent. |
| 00:18:50 | Dion | Okay, what's going on with you, Joe? |
| 00:18:53 | Dion | Just bear with me one moment and we'll get Joe up here. |
| 00:18:57 | Dion | There we go. |
| 00:18:59 | Dion | Okay, so we'll come back to Joe now. |
| 00:19:02 | Joe | Hi. My name is Doctor Joe, and I am the research director and advisor on the future of work here at Ibers. |
| 00:19:10 | Joe | Just a bit of background. Ibers is a research and advisory firm. |
| 00:19:14 | Joe | We do a lot of exploration of emerging technologies, and we help all of our clients who are typically the Typically the upper middle tier to higher end organizations in their IT planning. |
| 00:19:27 | Joe | Now. |
| 00:19:30 | Joe | Something always is amusing to me. You know, I'm really lucky. |
| 00:19:33 | Joe | I get to see what's coming up in the labs. I get to see technology before it hits the marketplace in many instances. |
| 00:19:40 | Joe | But more importantly, I get to see how it lands, how people are actually using the technology. |
| 00:19:45 | Joe | So, it's a great role to have. |
| 00:19:48 | Joe | But what amuses me is that we keep on hearing these phrases such as, you know, oh, this new thing has just happened. It's just exploded on the scene. |
| 00:19:57 | Joe | Technology is advancing at a pace that we can't imagine, and we've been hearing that for 30 years. I've been in this game a long time now. |
| 00:20:06 | Joe | Yes. |
| 00:20:07 | Joe | Byte by byte and storage by storage unit technology is increasing exponentially. |
| 00:20:15 | Joe | But that's not the same thing as saying that the technology is surprising. And it just it just comes down to the marketplace. |
| 00:20:22 | Joe | It explodes and that it's unpredictable. |
| 00:20:25 | Joe | Absolutely not. The march of technology is very predictable. |

| | | |
|---|---|---|
| 00:20:30 | Joe | And this includes people like Dion have been involved in in AI for decades. |
| 00:20:40 | Joe | And in many of the ways the technology is evolving, but it's evolving to a point. This is what we get confused about, you know, technology suddenly being in the marketplace. |
| 00:20:50 | Joe | At some point, we hit an inflection where the cost of using the technology, the cost of processing, whatever that is, is at a point where it's effectively consumerized and generative. AI is a perfect example of that. |
| 00:21:06 | Joe | Microsoft threw. |
| 00:21:10 | Joe | Billions of dollars into one of the AI ventures, which enabled it to deploy effectively ChatGPT, Committee, amongst others, at a cost that was extremely reasonable, and the cost of creating the AI models and the cost of deploying those AI models continues to decline. That doesn't mean that AI is cheap. |
| 00:21:32 | Joe | Now this situation tends to lead to overreactions in the market. People get extremely excited. |
| 00:21:38 | Joe | I remember when the iPhone hit the market, you know, it wasn't the first mobile phone. It certainly wasn't the first graphical mobile phone. |
| 00:21:45 | Joe | It was a damn good changed mobile phone, but its price point created an entirely new marketplace, and it was really the price of wireless data that drove that innovation more than the device itself. |
| 00:22:00 | Joe | There's many examples of this all through history. Now, when it comes to AI, we have to realize that over the last two years, we have been in the generative AI space equivalent of going back to using a computer With or without a mouse and with dots, you know, it's really clunky, this idea that you would have users prompt and interrogate these large language models manually is, quite frankly, ridiculous. |
| 00:22:30 | Joe | And in our projections of AI that we did almost 14 years ago, we expected that something like this would happen. |
| 00:22:39 | Joe | It's happened very aggressively. |
| 00:22:43 | Joe | So, what does the future hold? |
| 00:22:46 | Joe | The future, certainly. What many of the AI vendors are seeing at the moment is that their ChatGPT, like their copilot like interfaces, are landing well to a few people, but they're not expanding through the business. |
| 00:23:02 | Joe | And the reason for that is really simple. |

| | | |
|---|---|---|
| 00:23:05 | Joe | It's extremely hard to demonstrate true financial value when people are having to having to manually every single time they want to use the I manually work with it. |
| 00:23:19 | Joe | It's a fun experience, but it's not a productive experience. |
| 00:23:24 | Joe | And the result is very clearly there's a few results. Firstly, job vacancies for AI specialists and property specialists have plummeted. |
| 00:23:34 | Joe | So, organizations have figured yeah, it's not probably not the right thing to do. |
| 00:23:37 | Joe | Another is that Microsoft themselves are not able to leverage those big 300 contracts, you know, trial contracts. Proof of. |
| 00:23:49 | Joe | Proof of concept contracts with their enterprise clients for copilot. They've not been able to generally roll those out more broadly. |
| 00:23:58 | Joe | And as a result, they are actually experimenting with different licensing options already to try and drive copilot the copilot for 365 utilization. |
| 00:24:10 | Joe | There are many other examples of this at the same time. Ironically, the cost of many of the large language models continues to fall, and we get new entrants almost on a weekly basis. |
| 00:24:25 | Joe | You have to be very, very careful about the financial, the true financial machinations of those AI vendors where they're really getting their cash from. |
| 00:24:35 | Joe | But this is going to be is going to continue. It basically leaves us in a situation where a lot of organizations are saying AI is not. |
| 00:24:45 | Joe | It's too hard for us to show the value of the generative AI. The reason behind all of that is, as I said before, we're using these tools to the same extent, the same mindset as basically a computer without a mouse. |
| 00:25:03 | Joe | It's going to evolve very rapidly. |
| 00:25:05 | Joe | So, where are we heading now? |
| 00:25:08 | Joe | 2025 is going to be the year of AI agents. |
| 00:25:12 | Joe | Every man and his dog is going to be suddenly. Promoting AI agents. |
| 00:25:17 | Joe | And they're going to be talking about nondeterministic computer logic. And all this sort of stuff. |
| 00:25:23 | Joe | Sorry, it's not going to cut it. We're going to see and we are predicting the. |
| 00:25:27 | Joe | Same situation in 2025 early to mid 2026, where all of these AI agents are underperforming in the marketplace because at their heart. |

| | | |
|---|---|---|
| 00:25:41 | Joe | This idea of being non-deterministic means that it's very hard to pinpoint a true business. Value and productivity gain. |
| 00:25:49 | Joe | There will definitely be some really strong. Examples, but overall it's still missing the mark. |
| 00:25:57 | Joe | Now let's go back 14 years to when we started writing really extensively about AI. We positioned all of these advances in the realm of automation. |
| 00:26:12 | Joe | It's not sexy, but it's right. |
| 00:26:15 | Joe | You see, you can only demonstrate true business value when you automate something, when you remove work, or remove or streamline a process or remove a process entirely. |
| 00:26:27 | Joe | And machine learning and generative AI and graph. And all of these tools can be applied to processes, to workflows. |
| 00:26:39 | Joe | So, we believe that organizations that can orchestrate and understand that they need to orchestrate not just, you know, ChatGPT or some sort of challenge, but they orchestrate all of these different AI services, each of which has different strengths and weaknesses, each of which have different security profiles as well. When you can orchestrate that as part of your organization's processes, Since effectively making AI invisible to the average user. |
| 00:27:11 | Joe | No more manual prompting. Sorry, that should just disappear when you orchestrate it. |
| 00:27:17 | Joe | Suddenly it becomes very, very attractive to businesses. |
| 00:27:24 | Joe | Now I know the reason why they wanted me to speak today is because I've held that position for a while now, and the products which are rolling out are really in that in that zone. |
| 00:27:36 | Joe | So, I just want to say, you know, there is a reason I'm speaking here, but that doesn't take away from the fact that orchestration, that thinking about AI as nothing more than a service, it's not a product, it's a service that fits within some nominated area of your business that can be dramatically automated. |
| 00:28:01 | Joe | Now there's plenty when you start, when you flip your head around to that, you suddenly realize that there's many, many areas where AI can be applied very effectively. |
| 00:28:10 | Joe | Yes, you can look at transcription. Yes, you can look at all these other things. |
| 00:28:14 | Joe | But when you think about what else? |
| 00:28:18 | Joe | Customer service, how fun could you take that level of automation? |

| | | |
|---|---|---|
| 00:28:23 | Joe | You take a look at things such as planning or research. Even in our business, it's dramatically transforming our business because we are literally creating very, very complex orchestrations of AIS, multiple of them that effectively I would say eliminate but dramatically remove a lot of the complexities of doing research. |
| 00:28:45 | Joe | One orchestration that we have takes 47 to 50 minutes to run. |
| 00:28:50 | Joe | It consists of up to 1800 different AI calls iteratively working through material, and will eventually just give us an output that says, here's all the information globally known on this particular subject, and here's the answers that you've created. And this is not Google. |
| 00:29:07 | Joe | This is not perplexity. This is real research. |
| 00:29:11 | Joe | That was a task that used to take us two weeks to perform. It's now been shrunk down to an hour or so. |
| 00:29:17 | Joe | But of course the humans still in the loop on that. We just cut out the grunt work. |
| 00:29:21 | Joe | So, this idea of creating these highly effective orchestration chains will be the future. |
| 00:29:28 | Joe | I believe that the major vendors are still very much stuck on this idea of getting return on the investments for their, you know, prompting models, their GPT like services. |
| 00:29:39 | Joe | I think they're going to double down on trying to extend that through agents in the coming year, and it won't be until about 2026, mid 2027, where they'll actually figure out that these just aren't working in the marketplaces and platforms that allow you to stitch together multiple services, Whether they be from, you know, you as one big vendor say, you know, Microsoft wants to own it or Amazon wants to own all or whether they be independent platforms that allow you to bring in the best of breed and open source from wherever you want. I think those platforms will be the future. |
| 00:30:15 | Joe | So, I hope this has been informative. If if you do have questions, please reach out to me on LinkedIn. |
| 00:30:21 | Joe | I'll be more than happy to engage in conversations and fierce debate on this topic. I know that not everyone agrees with me on this, but I'm old enough and I'm crunchy enough and I've been seeing these same grand announcements. |
| 00:30:37 | Joe | Oh, this is moving so fast or this is revolutionary. I've seen them so many times to see the patterns in these. |
| 00:30:44 | Joe | And ultimately, when we do look at it, things like AI are very easy to predict their future from the technology perspective. |

| | | |
|---|---|---|
| 00:30:53 | Joe | What's harder is how and when organizations will use them. |
| 00:30:58 | Joe | Thank you very much and enjoy the rest of this, this discussion. |
| 00:31:04 | Joe | Hi. |
| 00:31:05 | Dion | Okay, so I'll hand back over to Phillip now, but I hope everyone found Joe's perspectives interesting. And I am one of the ones that argues and debates with him. |
| 00:31:13 | Dion | And we've been doing that for more than 20 years with each other. So, over to you, Phillip. |
| 00:31:17 | Philipp | Yeah. I want to take a bit of a kind of extended view of this concept of agents and where it's heading in the future. |
| 00:31:24 | Philipp | So, as Joe just mentioned, so we're talking about the orchestration of what we call agents, which are basically just AI empowered components that kind of have to collaborate and. Well, and as the word orchestration implies, this is typically done by kind of really arranging them in a certain way that they kind of communicate with each other, they kind of pass jobs through the workflow and so on. |
| 00:31:48 | Philipp | But when we talk about agents, we actually also talk about, yeah, how this process of arranging different components. |
| 00:31:58 | Philipp | It can be done and it doesn't have to be pre-planned. It could also be automatically generated. |
| 00:32:04 | Philipp | So, this plan, how these agents interact with each other might also be done with like basically a fully automatic planning system. So, this is like an old concept out of AI that if you have a complex problem that can be solved in a single step, that requires multiple components, that you first have to come up with a plan. |
| 00:32:24 | Philipp | And that kind of is a sequence of steps as illustrated here. And then you execute the first step. |
| 00:32:29 | Philipp | And then kind of a key idea here is after you execute the first step, you're actually going to replan. You're going to rethink how you're going to execute the next steps. |
| 00:32:37 | Philipp | Because maybe the first step failed. |
| 00:32:39 | Philipp | And therefore you have step two now is completely obsolete. So, you actually have to replan. |
| 00:32:43 | Philipp | So, this is kind of the old idea of kind of planning replanning. And this is something we see now happening with language models where you ask the language model first what is the sequence of steps to be taken, and then what is the first step. |
| 00:32:53 | Philipp | And step and then you go from there. |

| | | |
|---|---|---|
| 00:32:57 | Philipp | To give you like some really concrete examples where this is kind of the paradigm that has been pursued. So, this is kind of a just started DARPA program. |
| 00:33:04 | Philipp | So, from funded by the US government where the grant goal is to automatically verifying scientific claims. |
| 00:33:14 | Philipp | Of course, these claims are rather complex, and you have to kind of break it down in various components and all that has to be done automatically. So, you have to kind of take the claim, break it out into pieces, and then basically come up with which is called your reasoning chains or reasoning trees. |
| 00:33:29 | Philipp | That might involve more than just large language model interactions. They might also involve maybe even running small scale experiments or any other computational aspects to it. |
| 00:33:40 | Philipp | So, that's what we kind of have to kind of address there in the future. |
| 00:33:46 | Philipp | Just to give you one concrete example of such a method Is. Here's something we developed, developed by colleagues of mine at Johns Hopkins University with that exact process is brought down. |
| 00:34:00 | Philipp | So, here the idea is that you have a complex claim here. Something about plants and chlorophyl doesn't matter what it is, and you break it down into pieces and you ultimately prove each sub piece by linking it to a fact that is expressed in natural language. |
| 00:34:16 | Philipp | So, you imagine, take all of Wikipedia, write down all the facts, and that's kind of your fact base you work with. And any claim has to be broken down by kind of combination of inference rules and so on. |
| 00:34:27 | Philipp | But ultimately at the ground with natural language facts, you can see this kind of concept applies to various other domains like regulations, laws. That's kind of the fact that ultimately you have to drill down in a kind of legal judgment. |
| 00:34:40 | Philipp | And yeah, so the task is then, you know, you have to verify the state amendment. The solution is that you recursively decompose a statement to sub states, and ultimately you have to prove these against the facts. |
| 00:34:54 | Philipp | The interesting thing here is that the facts nowadays might come in all kinds of different ways. So, I just talked about text, but it might also be images. |
| 00:35:02 | Philipp | It might be audio, it might be databases, it might be charts, diagrams, tables. |
| 00:35:07 | Philipp | So, ultimately what a language model draws kind of evidence from to kind of ground. |

| | | |
|---|---|---|
| 00:35:16 | Philipp | This reasoning tree in facts might come down to really complex interactions with all kinds of multimodal knowledge. |
| 00:35:25 | Dion | All right. Let's get to one of my interesting topics. |
| 00:35:28 | Dion | So, Philip and I had a fierce debate over this topic. That's probably to say the least. |
| 00:35:35 | Dion | But in the end, I think we came to a general consensus. So, based on the information that's been out on the web and comments and many other things, we believe that meta this year is going to start charging for llama access. |
| 00:35:50 | Dion | There's many reasons for it, but one of the notable ones is the significant cost. Even though meta has a lot of money, it still takes a lot to train these. |
| 00:36:03 | Dion | You know, they're on record saying that they use 25,000 GPUs to train llama three and llama four. They're using 100,000 GPUs. |
| 00:36:13 | Dion | Now, these are 30,000 each. So, when you do the math, it's real significant large money. |
| 00:36:20 | Dion | But there's many other reasons. The biggest one is market diversification. |
| 00:36:26 | Dion | You know, meta is stuck in a world right now that everything revolves around advertising having another revenue stream. |
| 00:36:34 | Dion | And don't forget, I talked about these companies also building up into the application space. So, having the ability to generate margins just like Apple does in their app store from Metamodels makes a lot of sense. |
| 00:36:50 | Dion | So, the rising costs, you know, training llama three when you add up all the costs. So, the hardware alone, you know, was significant. |
| 00:36:59 | Dion | You know, but when you add up all the human resources, the electricity, getting the data ready and all that, it was about $1 billion. |
| 00:37:07 | Dion | And llama four is estimated, you know, just hardware costs alone at $3 billion. |
| 00:37:14 | Dion | Now Meta's got a lot of widespread adoption for their models. |
| 00:37:19 | Dion | In December last year, the downloads were going through the roof. A million downloads a day of llama 3.3. |
| 00:37:28 | Dion | It's extensive. |
| 00:37:29 | Dion | So, they've had 650 million downloads in total. |
| 00:37:34 | Dion | And the competitive landscape, you know, their competitors are starting to monetize in various ways too. So, you know, we don't expect |

| | | |
|---|---|---|
| | | that they're going to charge everything you know at high levels, or there's still going to be freemium access and various other things, but there will be different licenses. |
| 00:37:51 | Dion | They already charge organizations like AWS and some of the big companies like KPMG, substantial licensing fees. |
| 00:38:00 | Dion | So, we expect they're going to have a freemium model for noncommercial and smaller enterprises. |
| 00:38:07 | Dion | They're going to focus on mid to large size enterprises. And this is going to probably have a ripple effect across the industry, where people have to reevaluate their strategies and look at it all. |
| 00:38:19 | Dion | But it makes sense, you know, Meta's hybrid approach of different licensing models for different size business and at the same time looking at entering applications. And in the blog post, you'll see the kinds of applications that they could potentially go for and some of the other things that companies are specializing in already. |
| 00:38:39 | Dion | Okay, over to Philipp on the Rise of people first. I and rag. |
| 00:38:46 | Philipp | Yes. I'm only going to talk about what it actually means to deploy AI at the moment. |
| 00:38:50 | Philipp | So, yes, there's a foundation model, but then there's all this other stuff that actually makes it practical. So, there's retrieval augmented generation. |
| 00:38:58 | Philipp | There is knowledge graphs. We already talk about agents and of course the foundation models at its basis. |
| 00:39:06 | Philipp | So, all this put together will outperform kind of building an actual kind of fine tuning and foundation model. |
| 00:39:16 | Philipp | So, having a custom AI model in most business scenarios. So, you can actually build a lot around the foundation model. |
| 00:39:22 | Philipp | You don't actually have to adapt and improve the foundation model, which is a very, very costly endeavor. |
| 00:39:27 | Philipp | So, this has a lot of benefits. There's efficiency gains. |
| 00:39:30 | Philipp | So, this modular AI offers faster deployment reduce reputational overhead and cost efficient scalability. It also makes it much easier to adapt to changing needs on use cases. |
| 00:39:43 | Philipp | So, it combines real side insights, structured reasoning, modular orchestration to address dynamic business needs. |
| 00:39:49 | Philipp | And yeah, this all is kind of a new standard. Business will prefer modular systems for flexibility, scalability and real time performance over resource intensive custom models. |

| | | |
|---|---|---|
| 00:40:02 | Philipp | So, why is it that modular AI outperforms custom models? So, one is cost efficiency. |
| 00:40:09 | Philipp | So, the custom models require significant financial and technical resources. So, this basically means you have to take a very, very large model. |
| 00:40:17 | Philipp | And even more so when you use the model, when you train the model, you need even more computational resources and you basically have to do massive training on it. So, this is a very, very expensive thing to do. |
| 00:40:29 | Philipp | So, therefore it's better to kind of have kind of general models that then you build modular systems around. |
| 00:40:37 | Philipp | So, this is then. Yeah. |
| 00:40:40 | Philipp | Things like Rag as a method, knowledge graph as a method. So, these are all kind of then provide information to the model on demand. |
| 00:40:48 | Philipp | So, when you kind of interact with the model that information comes from outside sources and gets combined with a foundation model to kind of draw conclusions. So, that. |
| 00:41:00 | Philipp | Has all kinds of kind of real time relevance. So, you can actually update the models much more quickly. |
| 00:41:05 | Philipp | You can add more things in your database immediately, so your rag can dynamically retrieve real time data for up to date outputs, something that couldn't be in kind of the base foundation model because these get rarely updated. |
| 00:41:19 | Philipp | Also, the Knowledge Graph allows you to kind of have more structural and contextualized information for deeper reasoning. And all this makes it also easier to deploy because you have, you know, the pre-trained models on one component API calls to this or other kind of. |
| 00:41:37 | Philipp | The kind of logic units in the back. |
| 00:41:39 | Philipp | And then you can integrate that in various ways and workflows, and you can much quicker experiment with that. |
| 00:41:46 | Philipp | And also this enables much more interoperability. So, the key components for kind of really building AI systems today is basically combine these components. |
| 00:41:55 | Philipp | So, there's retrieval augmented generation where you kind of retrieve relevant information on the fly and kind of combine it with the abilities of the language model to reason over it. |
| 00:42:04 | Philipp | Knowledge graphs. You already mentioned AI foundation AI agents and foundation models. |

| | | |
|---|---|---|
| 00:42:09 | Philipp | So, we believe that the implication for that for industry is that module AI will dominate most use cases. Reducing dependency on custom solutions and custom models will remain relevant for specialized, highly regulated, regulated tasks. |
| 00:42:23 | Philipp | So, if you have a lot of training data for a particular task, actually machine translation is one of them. We'll talk about that more later. |
| 00:42:31 | Philipp | It actually does make sense to kind of build a custom model because the amount of training data you have for a particular task is still relevant. But for most practical use cases, having this kind of more modular approach with various components will be much more beneficial. |
| 00:42:50 | Philipp | Let me also kind of go a little bit more into like how these models are being used and what actually AI today means. |
| 00:42:59 | Philipp | So, when we actually look around here, what are really the use cases people talking about? So, we already talked about coding and apparently got 10% ten times faster coder. |
| 00:43:11 | Philipp | I have similar stories. I can't put a number on it. |
| 00:43:14 | Philipp | But yeah, it's definitely much more productive. |
| 00:43:17 | Philipp | And I hear a lot from people as writing assistants. And also already mentioned that you can make PowerPoint slides now with ChatGPT. |
| 00:43:24 | Philipp | So, these are kind of the use cases that are really compelling that people talk about a lot. |
| 00:43:31 | Philipp | So, and these are kind of the good use cases. They also kind of the misuse of this technology. |
| 00:43:36 | Philipp | So, it's kind of popular now for apparently students to cheat in the homework by just asking ChatGPT to do their homework for them and pass it off as their own. And there's like this famous case from, I think, about a year ago, where a New York lawyer kind of created a legal brief for ChatGPT. |
| 00:43:52 | Philipp | So, yes. |
| 00:43:54 | Philipp | So, this is kind of an interactive scenario where you, as a human, work on a task and the large language model or the AI helps you do the task, but ultimately you are still responsible for basically executing the task. |
| 00:44:13 | Philipp | So, in contrast to that, if you use large language models, that's fully automatic systems that bears a lot of risks. So, every AI system has an error rate. |
| 00:44:23 | Philipp | And what is the acceptable error rate? That doesn't matter depends heavily on the task. |

| | | |
|---|---|---|
| 00:44:28 | Philipp | So, we've worked on machine translation for now a better part of 2 or 3 decades. And we always had the mantra of it has to be good enough for purpose and good enough for purpose might differ. |
| 00:44:38 | Philipp | If you kind of use machine translation when you travel around for tourism versus, you know, translating a legal brief. And here, for instance, the example self-driving car where the bar is really, really high. |
| 00:44:49 | Philipp | A self-driving car that 99.9% of the time doesn't crash is not good enough. |
| 00:44:55 | Philipp | There's some famous cases where fully automated systems, like a chatbot for Air Canada made up refunds, and basically the airline ultimately had to honor it. |
| 00:45:06 | Philipp | So, we see much more advantages to using large language models as interactive tools for humans, where the user vets the responses and only uses what's correct, and then may also iterate over prompts and responses to get desired results. |
| 00:45:24 | Philipp | So, another reason this is now kind of ties back to this idea of rag and kind of evidence based kind of reasoning that a human user ultimately has to validate if what he gets back from the AI is trustworthy. |
| 00:45:41 | Philipp | And if you can link it to actual evidence, that's kind of a very key feature of kind of trustworthy kind of ensuring trust. Because if whatever statements are generated by the language models are backed up by actual, you know, human authored documents with known sources, then based on your trust in the sources, you can then trust the statements and believe in the output. |
| 00:46:09 | Philipp | Much better. |
| 00:46:12 | Dion | All right. |
| 00:46:14 | Dion | This is a rather tricky and sticky one. |
| 00:46:17 | Dion | Regulatory security and compliance issues. |
| 00:46:22 | Dion | So, the key drivers. |
| 00:46:23 | Dion | There's definitely a lot more stricter Deregulation. So, GDPR, the UAE act and various other data sovereignty laws. |
| 00:46:32 | Dion | That's driving a lot of things. And, you know, it's forcing a lot of people to be a lot more careful. |
| 00:46:39 | Dion | But it's moving systems in the right direction where it's going to drive AI strategies in the right direction for the right processes, keeping it secure. Now, we have noticed when we're dealing with our customers, we're having to deal a lot more with infosec departments. |

| | | |
|---|---|---|
| 00:46:57 | Dion | Explain to them what we do, how data is used. |
| 00:47:00 | Dion | If we're using bedrock for a customer, then why don't we have it on prem and we explain the costs and they can still choose to have it on prem, or they can use bedrock. |
| 00:47:10 | Dion | So, they've got those choices. |
| 00:47:12 | Dion | There's critical shifts now happening from reactive compliance. When they get caught doing something or something silly happens to proactive governance, the business is now prioritizing transparency, risk mitigation and overall consumer trust. |
| 00:47:29 | Dion | If they get caught breaching these violations, the reputational damage alone, let alone the other monetary monetary punishments are pretty shocking. |
| 00:47:42 | Dion | Compliance and security are basically no longer optional, but foundational for long term success in any of these regulated spaces. You must comply or you can't play. |
| 00:47:54 | Dion | Now, the risk of noncompliance. Well, let's look at a few metals. |
| 00:47:58 | Dion | Fined 1.2 billion. Not long ago Uber 290 million and LinkedIn 310 million. |
| 00:48:05 | Dion | It really hurts. These are real monetary penalties that are significant. |
| 00:48:11 | Dion | They can do reputational damage, operational damage. |
| 00:48:14 | Dion | And even somebody as big as meta doesn't want to blow 1.2 billion on nothing. |
| 00:48:21 | Dion | Then Overall, they're looking at different organizational deployments, more secure private cloud or on premise solutions to keep better control. Because every day in the media you see that you know, someone else has been breached, somebody's data has been lost, and hundreds of thousands of people's data is being lost on a daily basis. |
| 00:48:43 | Dion | So, there's strategic solutions. There's more advanced cyber security, localized infrastructures that I mentioned, federated learning, region specific data centers, for example, Amazon, they now have in a number of countries. |
| 00:48:58 | Dion | Amazon. Avnet. |
| 00:49:00 | Dion | So, these are government only systems that run aside from the main AWS the data privacy disconnected from the main AWS. |
| 00:49:11 | Dion | Okay. And that leads into the next point self hosted AI. |

| | | |
|---|---|---|
| 00:49:15 | Dion | So, on premises and private cloud there's going to be a resurgence for the exact reasons that I just talked about. So, you know, into, you know, in 2005, sorry, there's a typo there. |
| 00:49:29 | Dion | Self-hosted. I will start to reshape enterprise strategies. |
| 00:49:33 | Dion | A lot of countries are really hammering, you know, compliance on prem data sovereignty in particular. It's driven by control, customization and compliance needs overall. |
| 00:49:45 | Dion | And the regulatory demands are huge. |
| 00:49:48 | Dion | Now our platform language studio runs entirely on prem or in private cloud, and it can even be air gapped from the entire internet. So, that's one of our strategic advantages over all of our competitors. |
| 00:50:03 | Dion | But many people are just using regular cloud based AI. So, something as simple as Google Translate right now, if you copy and paste a company's contract into Google Translate, you're actually in breach of many of these regulations straight away. |
| 00:50:17 | Dion | And these things are going to start to get more strict. |
| 00:50:20 | Dion | Also, there's hybrid models where you can combine on prem and cloud integration or on prem and AI on demand, which is something like bedrock, which ensures data privacy all the way through. |
| 00:50:34 | Dion | These also allow for prototyping and especially scaling. |
| 00:50:38 | Dion | So, as you grow, rather than buying hardware, doing, you know, self-hosted but private cloud or AI on demand can be a lot better. |
| 00:50:50 | Dion | Okay, so I mentioned the different deployment models, their on premises private cloud and hybrid. |
| 00:50:56 | Dion | The big leading applications in this are healthcare, finance and government. But regular businesses are doing it all the time as well. |
| 00:51:03 | Dion | And there's a lot of advantages. Obviously your compliance, it gives you control of your data and tailored customization and scalability and flexibility. |
| 00:51:12 | Dion | So, it is well worth looking at these options. |
| 00:51:17 | Dion | If you need the data privacy, you want to access all of your data without having to worry about who's got it secure, who's holding the keys, who's hacked their system. Then self-hosted in private cloud or on premise is a pretty good option. |
| 00:51:33 | Dion | Okay, let's move on to the other half of the presentation. Now the language processing predictions. |

| | | |
|---|---|---|
| 00:51:41 | Dion | Okay, so this is an interesting one. It's something we get asked all the time when will computers beat humans? |
| 00:51:48 | Dion | Well it's happening in some areas. |
| 00:51:51 | Dion | And let's be very, very clear here. |
| 00:51:53 | Dion | There are multiple use cases we've seen where AI can match or outperform humans and it's increasing. |
| 00:52:02 | Dion | The humans are still superior in many areas, but AI is catching up and there's many areas that AI has an edge. |
| 00:52:10 | Dion | So, for example speed okay, you know, a human translates at 3000 words a day. |
| 00:52:17 | Dion | Just a single machine can do that in a couple of seconds. And bigger machines can do billions of words a day. |
| 00:52:23 | Dion | Scalability on repetitive tasks. Large scale tasks. |
| 00:52:27 | Dion | Accuracy. Yes, I said accuracy. |
| 00:52:29 | Dion | When things are done properly, you do the right terminology work and AI helps there too. I'm not just talking about translation here, by the way. |
| 00:52:39 | Dion | I'm talking about the entire localization stack, everything from transcription to onboarding data to project management translation. Certainly their quality control, terminology analysis and extraction and definition creating style guide using AI. |
| 00:53:01 | Dion | These are all human steps, right? But AI is stepping into these places already and it will fully automate many of them over time. |
| 00:53:08 | Dion | Not all of them. There will still be many areas for humans and in fact it will actually increase human activity in many areas. |
| 00:53:16 | Dion | So, automated quality assurance, you know, right now if you do QA, in many cases, unless you have a huge budget, you're doing spot checks. Automated QA can do QA on every single sentence. |
| 00:53:30 | Dion | You know, that's something that humans perhaps cannot do in a time frame or just because of a cost basis. |
| 00:53:35 | Dion | So, there's operational cost reasons to do it. |
| 00:53:39 | Dion | It's getting better enhanced reasoning and domain expertise and improved interfaces and integrations. |
| 00:53:46 | Dion | So, where is it superior now? |
| 00:53:48 | Dion | Well, bulk content handling, e-commerce catalogs, customer reviews, hotel reviews, those kinds of things. Real time updates, games, domain |

| | | |
|---|---|---|
| | | specific tasks actually, specifically law, medicine and tech are really doing well because of its ability to handle terminology so well. |
| 00:54:09 | Dion | But the terminology Work has to be done. |
| 00:54:13 | Dion | Now that's where you come to the next one. It's also a negative terminology. |
| 00:54:16 | Dion | Consistency across projects. Managing multilingual projects on large scale data sets can be challenging. |
| 00:54:24 | Dion | Also, the data sets it's learning from are often inconsistent. So, some of our new tools with our automation tools will give you an opportunity to normalize terminology. |
| 00:54:34 | Dion | So, you know, we might talk cloud computing today, but a few years ago they might have called something very similar client server or a variation of client server. So, the terminology and vocabulary is evolving. |
| 00:54:47 | Dion | The advantages over human workflow across speed, cost efficiency and error reduction. Is it a magic wand that solves everything? |
| 00:54:55 | Dion | Of course not. But the point is that there are more and more use cases where machine or machine plus human can easily exceed human only the last points. |
| 00:55:08 | Dion | It's all about collaboration. |
| 00:55:10 | Dion | It's not about just this machine running off and doing all the work. |
| 00:55:14 | Dion | It is very much about collaboration. And there'll be new roles for humans. |
| 00:55:19 | Dion | There'll be roles that humans are can still excel in. For a long time to come. |
| 00:55:23 | Dion | But we'll talk about one of them a little bit later. Creativity and emotional depth in literary marketing and poetry and things like that. |
| 00:55:32 | Dion | Handling ambiguity. The AI is getting better, but it still needs time to mature. |
| 00:55:38 | Dion | So, the best approach going forward is a collaborative approach with Post-editing to get polished results, leveraging AI to expand reach, and while maintaining quality and scalability without having to sacrifice cultural and contextual relevance. |
| 00:55:54 | Dion | Okay, I'm going to hand back to Philip. |
| 00:55:56 | Dion | Yeah. |
| 00:55:56 | Philipp | Now we're going to talk about like how AI changed machine translation, although it's a bit of a weird way to talk about it, because if |

| | | |
|---|---|---|
| | | you really look at AI today, you talk about the transformer model, which is a machine translation model. So, just to be for clarity when I say now. |
| 00:56:09 | Philipp | Machine translation, I talk about a dedicated machine translation system that was built to. Do translation. |
| 00:56:15 | Philipp | And when I talk about I mainly mean large language models or related. Technologies that have a much broader view at language or even other modalities of data. |
| 00:56:24 | Philipp | So, we believe that AI is is the translation. So, where we not just use a dedicated machine. |
| 00:56:29 | Philipp | Translation component, but we also build kind of a very complex environment of AI around it will. Become an industry standard in 2025. |
| 00:56:39 | Philipp | And I will extend beyond just the kind of. Machine translation part of this. |
| 00:56:44 | Philipp | But the entire workflow of localization. So, that allows humans. |
| 00:56:49 | Philipp | Translators to focus on creative culturally and and kind of more sensitive, nuanced tasks. |
| 00:56:57 | Philipp | This is kind of give you a bit of a picture of how this looks like. So, this is. |
| 00:57:00 | Philipp | Kind of a typical translation workflow where you have a document you translate, you send it to a machine translation, you have translation of the MT output and the human post editor and human quality control and then delivery. So, this is kind of where we have been maybe already for kind of maybe a better half of a decade. |
| 00:57:17 | Philipp | So, you get a lot of productivity gains out of MT because you only have to fix the mistakes, which is much faster than writing it from scratch. So, let's just say here, if you work on MT, you're going to be 5,050% faster. |
| 00:57:33 | Philipp | So, now as we're going to talk about is an additional component of automated AI Post-editing. So, this is something we currently spend a lot of investment on how to get that right and how to kind of combine the raw MT output with very dedicated automatic revision components, where now the AI that's doing revision. |
| 00:57:54 | Philipp | Okay. |
| 00:57:55 | Philipp | The next step would be then also to put in automatic quality control. So, here you basically have the ability to say okay at this point we're pretty sure the translation is good enough already, and there's nothing wrong with it, so no human poster even has to look at it. |

| | | |
|---|---|---|
| 00:58:10 | Philipp | So, that kind of cuts out even the need of a human post editor to look at the translation. And yeah. |
| 00:58:19 | Philipp | So, this goes all the way back to kind of the human quality control. So, after the human post editor, there's typically a human revision stage, and even we might be able to skip that step as well. |
| 00:58:30 | Philipp | So, that doesn't have to be done either. |
| 00:58:33 | Philipp | Okay. |
| 00:58:34 | Philipp | So, by basically having evidence of things being all right and having feedback from from human quality control, this is actually a component that can improve over time and kind of bypass human through the entire process. |
| 00:58:50 | Philipp | So, this is kind of a rough outline where like two parts like we expect AI to make major contributions, but it's really the entire chain of localization. I'm not going to go through the list here. |
| 00:59:01 | Philipp | But if you just think about like from the beginning, from the initial consultation and kind of preparing research and, and and delivery and feedback from client, all these components kind of just ask for having improved kind of processes, aided by AI, since we're talking about AI and machine translation and kind of make this distinction between like the core component and the AI around it. |
| 00:59:32 | Philipp | We still kind of struggling a bit with that. The core team component can be very efficient, can be very fast, can be very cheap, and can be very deployable and adaptable. |
| 00:59:41 | Philipp | While the AI component is really, really expensive because this now involves these gigantic models and all the interactions for this. But there's increasingly interest or maybe can we just fold that all into the dedicated MT engine so we can still have a nimble, fast engine that does all these complicated things that we just talked about or also illustrated here at the start of the slide here. |
| 01:00:03 | Philipp | You know what you can do with large language models. You can specify style, you can provide terminology, you can add all kinds of constraints. |
| 01:00:10 | Philipp | These are all things you can do with large language models. And ideally you want to do this inside the machine translation model as well. |
| 01:00:18 | Philipp | So, just to kind of point to like ongoing research. So, this is work at Microsoft where they basically worked on kind of instruction, fine tuning machine translation models. |
| 01:00:28 | Philipp | So, we basically take a machine translation model. And now we also allow it to be fine tuned with instructions. |

| | | |
|---|---|---|
| 01:00:34 | Philipp | So, you not only say, here's a source sentence, but you provide the source sentence and additional instructions to then kind of guide the process of translation. So, this is really interesting work. |
| 01:00:46 | Philipp | So, here's a bit of a cartoon illustration of this. So, we typically have the trade off between in the lower left corner the MT which is not as capable in terms of kind of the style adaptation and terminology, coherence and so on, but it's fairly cheap. |
| 01:01:03 | Philipp | And on the top right you have the large language model, which is very expensive but has all these capabilities. And we basically want to look into how can we extend the MT models to have all these capabilities, while still being nimble and cheap, to be used as still a very small memory footprint and compute cost? |
| 01:01:21 | Philipp | I'm going now to the flip side, which is a large language models. How can we make them be better at translation? |
| 01:01:28 | Philipp | So, it's actually interesting. For several years people just kind of looked at how good can ChatGPT do machine translation? |
| 01:01:34 | Philipp | Although ChatGPT was not specifically optimized for machine translation, clearly it has seen a lot of translated data, so that's why it could do the task at all. But really, if you want to use a large language model for machine translation, you should actually train it to the task. |
| 01:01:48 | Philipp | So, clearly large language models are instruction tuned for various tasks. And one of these tasks probably should be machine translation, and it's relatively straightforward to do this. |
| 01:01:57 | Philipp | You basically take all the parallel data you have and convert it into chart format. So, here's a typical kind of instruction prompt there. |
| 01:02:03 | Philipp | Translate the following sentence from German to English and German. That's how this goes in English. |
| 01:02:07 | Philipp | This house is the house is big. And that's kind of the instruction data that you find in your model on. |
| 01:02:13 | Philipp | I just want to highlight two particular kind of instances of work where this has been recently done. So, this is work by Alma project. |
| 01:02:21 | Philipp | Sorry. A project called Alma is work done at Microsoft and also with students at Johns Hopkins being involved, where the important insights was that it's actually fairly important to pull this off by adopting a model like llama to to a machine translation that you also pre-train on a lot of monolingual text, so it's just much more aware of the text. |
| 01:02:42 | Philipp | And then you might not even need a lot of parallel data. So, in this paper, they really made the point that you need relatively little high |

| | | |
|---|---|---|
| | | quality parallel text to get really good translation performance, that kind of beat GPT 3.5. |
| 01:02:58 | Philipp | Here's another effort along those lines, which is from Unbabel, who put a lot of effort also to adapt language Llama to with large amounts of monolingual data and parallel data with various filtering with a kind of comic kind of filtering technology, and also instruction tuned data from other tasks and also multi-shot translation and, and interestingly, that both the training data and some of the models are publicly released. And you can use this. |
| 01:03:29 | Philipp | So, this is probably at this point the most serious effort actually to adapt a language model to machine translation. But it's actually interesting that the computational cost for that is actually not that massive. |
| 01:03:41 | Philipp | So, we do expect this to happen more and more here. The last thing I want to say about like this whole debate MT versus AI. |
| 01:03:48 | Philipp | So, these are results from like two months ago from the WMT competition, where basically a blind test set is. Presented to various online MT systems and large language models. |
| 01:04:04 | Philipp | But. The main purpose of this is to kind of for research teams to compete and. |
| 01:04:08 | Philipp | Try out new things. |
| 01:04:10 | Philipp | But at the end of the day, we're going to have a ranking. Where we basically see which performs better on a previously unseen test set that no other system had access to. |
| 01:04:18 | Philipp | And well, on the one hand, the good news is that the humans still win most of the time, and otherwise most of the time. And the reason why they don't always win is like, yes, if you hire translators to do a translation job, they might do work that is good enough for their purposes, but not the best they could possibly do otherwise. |
| 01:04:40 | Philipp | It's I mean, if you look at the blue boxes, these are all the dedicated MT systems and the green boxes, which are large language models. It's kind of a pretty open competition right now. |
| 01:04:49 | Philipp | it's not very clear what is really the best thing to do, and it really heavily depends on the language pair. |
| 01:04:55 | Philipp | Okay. Just finally, an outlook of kind of this debate of AI and MMT. |
| 01:05:01 | Philipp | So, the key impacts of AI integration interpretation is yes, it allows you to kind of automate repetitive and data driven tasks with efficiency. It enhances quality assurance even for lower budget projects. |

| | | |
|---|---|---|
| 01:05:12 | Philipp | It supports every translation stage, as I kind of pointed out, preparation, translation, editing and so on and so on. |
| 01:05:18 | Philipp | And there's a lot of benefits and future implications. So, it kind of makes everything cheaper, makes everything go better, improves global communication through efficient, accurate translations, ensures cultural relevance and nuanced understanding with AI human synergy, and redefines the translator roles as editors, mediators, and project managers. |
| 01:05:39 | Philipp | Okay. |
| 01:05:40 | Dion | Now we're going to get into an interesting part. |
| 01:05:44 | Dion | And in this part, I'm just going to Jump over to something on the other side. And I'd like to introduce you to a little friend of mine. |
| 01:05:55 | Dion | His name is Humphrey. |
| 01:05:59 | Dion | So, good evening, Humphrey. How are you today? |
| 01:06:03 | HUMPHREY BOT | Good evening. I'm doing well. |
| 01:06:05 | HUMPHREY BOT | Thanks for asking. |
| 01:06:07 | Dion | Okay, so. |
| 01:06:08 | HUMPHREY BOT | How about that? |
| 01:06:09 | Dion | So, today we're going to do a little bit of talking about a couple of topics. |
| 01:06:14 | Dion | Can you go and do a little bit of research on a company called Morgan Stanley for me. Just tell me three bullet points. |
| 01:06:19 | Dion | What do they do? |
| 01:06:23 | HUMPHREY BOT | I'll look into that for you. Hold tight. |
| 01:06:30 | HUMPHREY BOT | All right. Here's a quick overview of Morgan Stanley for you. |
| 01:06:33 | HUMPHREY BOT | They're a big player. |
| 01:06:37 | Dion | Okay. This happens sometimes when I'm running teams. |
| 01:06:39 | Dion | Unfortunately, it doesn't run so well. |
| 01:06:43 | Dion | Okay, Humphrey, we'll come back to you, and I'll put a recording on the web for this one to demo it better. |
| 01:06:49 | Dion | Anyway, that's what happens when you try to run all these tools on the same machine and at the same time. But let's go and look at how these are used. |
| 01:06:59 | Dion | So, conversational AI is really kicking off. |

| | | |
|---|---|---|
| 01:07:03 | Dion | We've got a range of offerings in that already. |
| 01:07:05 | Dion | And we're working in with partners to do a whole range of other features. |
| 01:07:11 | Dion | It's pretty amazing when you get these things going and I'll tell you how good they are. We did a fake customer support call where we asked the tool to imitate an agent for their company and research them online, and that's what I was going to do just then for Morgan Stanley. |
| 01:07:28 | Dion | And it's all live and in real time. It then went up to their management. |
| 01:07:31 | Dion | We did a recording and they thought the recording was real, and somebody had hacked their system and got access to internal data. That's how realistic it was. |
| 01:07:41 | Dion | And they were very concerned. But then they were pleasantly relieved when they worked out that the machine got everything it needed to do to act as a customer support agent directly off the internet and just their reference data. |
| 01:07:52 | Dion | So, real time translation and conversational AI are going to be omnipresent by the end of the year. It's just built into so many things. |
| 01:08:02 | Dion | Seamless communications across languages, integration into smartphone glasses, telephones or computers like I'm running now. |
| 01:08:12 | Dion | It's going to have a massive impact when the user interface changes from text and typing to talking and talking to almost anything, right? Talk on the telephone. |
| 01:08:26 | Dion | Now, you might have noticed that ChatGPT has just launched a new service. You can call one 800 ChatGPT and just talk to them on the phone, and it gives you an answer, just like I was doing then. |
| 01:08:36 | Dion | Okay. That simple. |
| 01:08:38 | Dion | So, you know, you might not have access to the web every time. |
| 01:08:42 | Dion | Customer assistance. |
| 01:08:44 | Dion | We had a proposal out to one of our customers where it was a large government department equivalent to the FBI. They had to do 20,000 interviews, preliminary interviews for a legal case. |
| 01:08:56 | Dion | And we put forward using these tools, and we're going down that path for future projects. We didn't have enough time in their timeline, but actually having the I have live interviews with victims, 20,000 victims that just wouldn't have been possible. |
| 01:09:12 | Dion | They were only going to interview 100 of them, and this made it possible to do 20,000, which they're planning to do in the same scenario, should they come up again. |

| | | |
|---|---|---|
| 01:09:21 | Dion | So, key industries are government, of course, healthcare, customer service, all these kinds of things. But it's not just multilingual assistance. |
| 01:09:29 | Dion | It can really hook into everything from your custom agents that go out and do the things like Philip talked about earlier. |
| 01:09:36 | Dion | I can just ask it to go and book me a me a haircut appointment, or book a restaurant for a visiting guest that I have tonight, or change my flight details, all these kinds of things, or look up things in the middle of a meeting, say, hey, my colleague said something ten minutes ago about the price of these widgets. What did he say? |
| 01:09:55 | Dion | Remind me. I've forgotten. |
| 01:09:57 | Dion | Oh, and now, at the end of the meeting, please tell me a summary of the key things that were talked about and drop them in an email to me. All of these things are immediately possible today, and we're shipping these products right now. |
| 01:10:08 | Dion | Expect to see a lot of competition in this space, but it's an amazing space to play in. |
| 01:10:14 | Dion | They're going to be on consumer devices. They're going to be inside teams. |
| 01:10:17 | Dion | Where in something like this I could just recall live captions, note generation and not just basic notes. The stuff that teams does is very, very simple and basic. |
| 01:10:28 | Dion | If you saw our webinar a couple of months ago with our business AI tools, you can see really comprehensive notes, breakdowns, even fact checking. So, I went to a conference in Dubai and Boris Johnson was on stage and I was in the back of the room. |
| 01:10:42 | Dion | I just recorded his presentation on my phone with 800 people around me. I got back to the office. |
| 01:10:47 | Dion | I said, right, analyze this, write it up. |
| 01:10:50 | Dion | And it went through. And it did X, Y, and Z and give me some great summaries quotes on different topics. |
| 01:10:56 | Dion | But at the end it says, here's two facts that he claimed that are lies and here's why they're lies right in the middle of the document. So, all of these kinds of things are possible public service announcements, all these kinds of things. |
| 01:11:09 | Dion | And it's multi-modal audio, text and visual inputs all together, hands free while you're driving. |
| 01:11:15 | Dion | Right. These kinds of things, all of this is possible okay. |
| 01:11:20 | Dion | So, it's fluid. It's natural. |

| | | |
|---|---|---|
| 01:11:23 | Dion | I'll post a couple of recorded examples of it on the on the website afterwards so that you can play them and see what's really possible. It really is truly amazing. |
| 01:11:33 | Dion | And now we're tying that into things on the back end so that you know it can do customer support. It can look things up. |
| 01:11:40 | Dion | It could register your product or register a complaint. All of these kinds of things. |
| 01:11:46 | Philipp | Yeah. Just want to give you a quick kind of glance at like what is involved in kind of this technology by combining kind of previously textual translation with machine translation with speech recognition. |
| 01:11:59 | Philipp | So, we basically have this task of someone directly speaking, and we basically want to translate it into another language. And it typically requires the steps of a speech recognition converting the speech into text then text translation. |
| 01:12:12 | Philipp | So, machine translation and speech synthesis. |
| 01:12:15 | Philipp | So, there's a lot of work on this. This is kind of a most exciting topic generally also building multimodal foundational models. |
| 01:12:22 | Philipp | So, this is kind of a really interesting space we're currently working in. So, the kind of really big goal in a lot of these efforts is to kind of convert this step by step process, process what's also called a cascaded approach to an end to end process. |
| 01:12:37 | Philipp | However. |
| 01:12:40 | Philipp | Yeah. So, yeah, let me just skip that. |
| 01:12:43 | Philipp | So, this also enables, for instance, that if you do this process live, you don't have to kind of go through all these steps. And you can do much, much faster kind of simultaneous translation. |
| 01:12:53 | Philipp | So, you don't have to kind of first wait for the speech recognition to kick in, and then for the text translation to kick in and then for the synthesis to kick in. That all adds to the delay. |
| 01:13:04 | Philipp | So, if you have to have an end to end system, you can do this much more quickly. So, this does involve kind of algorithms that figure out how much do you have to wait for input signal to come in before you produce output signals. |
| 01:13:18 | Philipp | So, there's like for instance, these kind of weight k policies that are being used. |
| 01:13:23 | Philipp | There's fundamentally the challenge that for all these kind of separate steps, we have either moderate or a lot of training data. But for the end to end process we process. |

| | | |
|---|---|---|
| 01:13:31 | Philipp | We have very little data. So, there's various things that can be done about it. |
| 01:13:35 | Philipp | So, one of the solutions is to kind of synthesize the pieces of the data that you don't have. So, if you only have a lot of translated text, you can synthesize the audio on the source or the target side. |
| 01:13:45 | Philipp | Or the other big solution is to kind of build these models incrementally. First start with the text translation model and then just add audio input modalities at the beginning and then synthesis at the end, and then kind of only at the end with the kind of the precious end to end data to fine tune it on, on the task. |
| 01:14:01 | Philipp | The other thing I want to talk about is also that we don't have to stop at speech. We can also kind of go all the way to video, because communication is not just the literal meaning of that can be reduced to text. |
| 01:14:13 | Philipp | There's a lot of emotional and nuance in the speech and the and the physical expression. So, basically we want to build models that also draw on video recordings, on the facial expressions, on the mouth movements. |
| 01:14:27 | Philipp | Just to kind of illustrate this here. So, you can imagine that in the input video, you get a lot of insight from the facial expression, how something is actually intended. |
| 01:14:37 | Philipp | And you can actually then take that into account when you do the translation, which also if ultimately, if you want to generate the video at the end, you also have a lot of challenges to basically map the output image to exactly the correct emotional kind of facial expression and gestures in the output video. So, these are kind of all really, really interesting and challenging problems. |
| 01:15:01 | Philipp | Here's for instance one example of how this works. Basically, you kind of start with a text to text translation model, and then you kind of add all this additional modality information on the source and target, kind of bring all these different means of communication by, by vision, by by speech, by text on the same kind of representation space to basically then build kind of an end to end system that can take both text, video and audio as input. |
| 01:15:30 | Dion | Okay. And the last topic today I believe, I think not 100% sure we'll find out in a moment, is about widespread use of AI for creative translation and localization. |
| 01:15:43 | Dion | So, you know, I can't think how long it's been, but for years, decades, people have said, I'm creative. A machine can never do what I do. |
| 01:15:53 | Dion | And I've heard that for decades. |

| | | |
|---|---|---|
| 01:15:55 | Dion | Okay, well, the reality is that machines are becoming creative, and in some cases they're even more creative than humans. I've been absolutely amazed at some of the graphics, some of the way things have helped me in my writing and things like that. |
| 01:16:10 | Dion | And so on. But combining, you know, a range of things now, it's usually a collaborative creation where it's a iterative. |
| 01:16:20 | Dion | Here's an image. Yeah, but I don't like that fine tune that add a seagull in the in the top corner. |
| 01:16:26 | Dion | Take out this tone. The brightness down on on this part of the image. |
| 01:16:30 | Dion | Or add shadows and off it goes. But it's doing creative tasks. |
| 01:16:34 | Dion | It's doing that in text mode as well. And it's doing it in a range of other things. |
| 01:16:39 | Dion | And when you hear some of the conversational AI, you'll just see how creative the voices get, where they're showing expression, you know, they're showing all sorts of things, and it becomes very hard to work out what's real and what's not. |
| 01:16:53 | Dion | So, it's combining the accuracy there. |
| 01:16:57 | Dion | So, you know it's going to change literature and even translation. It can be really creative. |
| 01:17:03 | Dion | So, let's look at some of the drivers. Obviously transcreation taking in intent and tone, cultural adaptation, understanding things. |
| 01:17:12 | Dion | They're specialized AI models are helping that a lot. But even regular llms can do a lot of it with the right guidance. |
| 01:17:21 | Dion | Human AI collaboration is the key behind the whole thing. |
| 01:17:25 | Dion | It's storytelling with a balance of efficiency and artistry together. |
| 01:17:30 | Dion | But let's have a look at some. |
| 01:17:32 | Dion | So, on the left here is Spanish and this is using our tape tools. |
| 01:17:37 | Dion | That's the original Spanish. And here's raw machine translation. |
| 01:17:41 | Dion | Okay. Now it's a perfectly good translation all the way down. |
| 01:17:44 | Dion | There's nothing wrong with it at all, except it's boring, even though it's a perfect translation. Now, the use case in this case was to make it for a telenovela being translated from Spanish into English, and they wanted it in synopsis form so that you could put it up and compel people by reading the synopsis to watch the episode. |

| | | |
|---|---|---|
| 01:18:07 | Dion | So, if I go here and say, you know, Fabio argues with his wife Carla, and she leaves the house, doesn't this sound better? Fabio gets into an argument with his wife, Carla, and she walks out on him. |
| 01:18:18 | Dion | In the morning. He has to take the children to school. |
| 01:18:21 | Dion | In the morning. He's left to take the kids to school on his own. |
| 01:18:25 | Dion | But this one's a real killer. Later, Carla comes home and asks him for a divorce. |
| 01:18:30 | Dion | Later, Carla returns and drops a bombshell. She wants a divorce. |
| 01:18:33 | Dion | She's out of there, right? So, it's that kind of creativity and that's guided. |
| 01:18:40 | Dion | Now. It's also able to change a lot of our other content to be richer. |
| 01:18:45 | Dion | So, here you can see Nacho takes the DNA test with the help from Emma, his secretary. And it's the same on both sides. |
| 01:18:52 | Dion | But when they want to do it to Johnny, which is exactly what's written in Spanish, they discover he ran away. But the translation that's been fine tuned. |
| 01:19:01 | Dion | When they want to administer the test to Johnny, it sounds so much better, right? |
| 01:19:06 | Dion | So, it's stylizing and making it suitable for purpose. And these are tools that we're getting a lot of fantastic feedback from. |
| 01:19:14 | Dion | Okay. |
| 01:19:15 | Dion | We can optimize to different writing styles. We used to do this by training different MT engines. |
| 01:19:21 | Dion | We don't do that anymore. We can just tune the outputs. |
| 01:19:24 | Dion | And Philip mentioned having this actually built into the engine itself. Now, where LM functionality and guidance is being built into the back end of the translation process directly in the NMT engines. |
| 01:19:37 | Dion | Now this is where you might be really surprised. |
| 01:19:41 | Dion | So, this is very thick, strong, nasty Australian surfer vernacular. And I'm from Down Under. |
| 01:19:49 | Dion | I'm from New Zealand and I can't understand this. But it's really hard. |
| 01:19:54 | Dion | You know you had a proper surf. You know, you get the sand all up your clacker and down your gobbler. |
| 01:20:00 | Dion | You know, all of these kinds of things. |
| 01:20:02 | Dion | No idea what they're talking about. |

| | | |
|---|---|---|
| 01:20:04 | Dion | So, our AI tools did analysis. Two friends lightly male. |
| 01:20:09 | Dion | Okay. You can see sharing a casual conversation. |
| 01:20:11 | Dion | What's the scenery? What's the situation then what's the tone? |
| 01:20:17 | Dion | So, it's pulling all of this out of just a handful of sentences. |
| 01:20:21 | Dion | Key elements that it can find. So, because of the vocabulary that was used, they've decided it's Australia, UK or New Zealand. |
| 01:20:30 | Dion | And that's accurate. |
| 01:20:32 | Dion | It's full of colloquialisms. Now what do these words mean. |
| 01:20:35 | Dion | So, cheers mate. Okay. |
| 01:20:37 | Dion | Simple enough. A quickie A quickie in this context is a brief snack or a drink or a meal. |
| 01:20:43 | Dion | A proper surf is a great or excellent surf. Clacker is your mouth okay? |
| 01:20:50 | Dion | Gobbler is your Adam's apple. |
| 01:20:52 | Dion | Beef Kap is slang for skull. An anteater is your nose. |
| 01:20:58 | Dion | Okay, so now that you've got definitions, you can translate that and you can choose the right slang or whatever. And you can even have the AI suggest the right terminology in the target language. |
| 01:21:09 | Dion | Okay. So, transcreation with these kinds of things, with idioms, with metaphors, emotional resonance are all their marketing campaigns that are tailored to local cultural context can be done today. |
| 01:21:23 | Dion | Okay, you can upload an existing marketing campaign and say use the same style, mirror this kind of context. |
| 01:21:31 | Dion | Okay. It deals with cultural missteps, even poetry. |
| 01:21:37 | Dion | Okay. Context aware. |
| 01:21:39 | Dion | Multi-Modal from text, audio and visual all the way through. |
| 01:21:44 | Dion | Okay. Now, it should be collaborative because it doesn't know everything, but it's really helpful. |
| 01:21:50 | Dion | And then workflow optimizations and hyper personalized marketing. Like I mentioned earlier, I was on Instagram and a video ad came up with me in the video directly. |
| 01:22:01 | Dion | So, all of these things are there. |
| 01:22:03 | Dion | And I knew there was one more. |
| 01:22:05 | Dion | Okay. The rise of no human in the loop translation. |
| 01:22:10 | Dion | No human anywhere. |

| | | |
|---|---|---|
| 01:22:12 | Dion | Well, maybe a little bit. Let's see. |
| 01:22:15 | Dion | Okay, so we have no human in the loop. What does that mean? |
| 01:22:21 | Dion | That means seamlessly operating without human intervention. |
| 01:22:25 | Dion | Okay. Fast, accurate, cost effective machine translation in some tasks that we're doing now is surpassing human levels. |
| 01:22:34 | Dion | Okay. So, we're currently doing a project I'll talk about in a moment that many of the translations to English are beating what a human translator would do, because it's taking in that creativity process. |
| 01:22:45 | Dion | It's a lot slower to translate than standard machine translation, but a lot faster than humans. But it's absorbing them and bringing them in, and then it's analyzing. |
| 01:22:55 | Dion | It's from Germany and it's this context and it's soaps and dramatic and it's applying that into the translation directly. So, it's taking first of all, it's doing speech recognition, then it's doing translation and then it's doing subtitles. |
| 01:23:10 | Dion | And I'll show you a little bit more about that in a moment. There's going to be big impact on business with this. |
| 01:23:16 | Dion | Both real time and batch mode. It really reduces the reliance on humans, which in many cases can be a good thing. |
| 01:23:23 | Dion | If you need an instant transcription, you know you don't want to be waiting for a human. Okay, so this is one of our customers, UFA in Germany. |
| 01:23:32 | Dion | They're processing about 20 or 30,000 hours of content. Okay, so it's a world first. |
| 01:23:38 | Dion | We launched it with them just a few months ago. |
| 01:23:41 | Dion | It's taking German content that's about 20 to 30 years old, up to just about eight years ago, something like that, maybe five years ago. It transcribes it with a custom transcription model that we fine tuned. |
| 01:23:52 | Dion | We then do speech automated post-editing that fixes bad speech recognition and fine tunes the structure, because speech recognition doesn't always give you the best sentence boundaries, things like that. |
| 01:24:05 | Dion | Then it does a quality control pass. Now some of the content is bad. |
| 01:24:09 | Dion | It's just really old and the audio quality is bad, so it gets flagged and a human can come in and check it. |
| 01:24:15 | Dion | If it passes automatically, it goes down to translate and then it translates and we use translation automated post-editing like I showed |

| | | |
|---|---|---|
| | | you with the synopsis that fine tunes it for soaps Sensationalizes it makes it dramatic and emotional and all these kinds of things. |
| 01:24:32 | Dion | Then it goes to quality assurance again, and if it passes there, then it goes directly to YouTube. No human in the loop. |
| 01:24:40 | Dion | Thousands of videos. |
| 01:24:42 | Dion | Okay. That's substantial. |
| 01:24:44 | Dion | Now, what this is about is good enough. |
| 01:24:48 | Dion | Good enough. Translation. |
| 01:24:50 | Dion | Now, is it perfect? No. |
| 01:24:51 | Dion | It picks up the music and and transcribes their theme song. It picks up a television in the background sometimes and transcribes that. |
| 01:25:00 | Dion | But compared to paying $1,000 versus less than $100 per episode. |
| 01:25:06 | Dion | Right? These are the are the differences. |
| 01:25:08 | Dion | It makes these use cases viable. It makes it possible to monetize them and drive things forward. |
| 01:25:14 | Dion | So, you know, many use cases, you can have no human in the loop because it's good enough. You can accept a non-perfect, but pretty high quality result right, as a trade off for mass, volume and low cost. |
| 01:25:32 | Dion | Okay, so there's many monetization opportunities that can unlock audiences. Films and media revitalizes old content, real time content, translation, translation, quality improvements, and all sorts of ways to give new ways to get market access to foreign markets. |
| 01:25:51 | Dion | Okay. And that's the last one. |
| 01:25:53 | Dion | So, if you have any questions, please feel free to jump on. We're almost at the end of our time, but we'll go a little bit over to answer some of the questions. |
| 01:26:02 | Dion | And I'm just going to roll up the questions now and see what we've got already. |
| 01:26:07 | Dion | So, Philip, how about this one? |
| 01:26:09 | Dion | What do you think on this? |
| 01:26:13 | Dion | Okay. How is text recognition coming along? |
| 01:26:17 | Dion | We have a large set of data confined to PDFs, and it would do wonders if these could be accessed for predictive purposes. |
| 01:26:25 | Philipp | Yeah, there's clearly a lot of work on that. So, this is kind of OCR and PDF are historically nasty. |

| | | |
|---|---|---|
| 01:26:34 | Philipp | So, this is something where actually kind of these AI vision models, text models have made significant progress. |
| 01:26:42 | Philipp | So, just to give you an anecdote, what I tried recently is basically coming out from scientific papers. Wanted to extract the tables and then they basically just images. |
| 01:26:52 | Philipp | And how do you get the text out and convert them into kind of readable format and just kind of tried out ChatGPT for that. And it actually worked really, really well. |
| 01:27:00 | Philipp | So, basically by combining kind of text models, image models in one big kind of component, which then not only can draw on what it can, just like identify from the characters on the screen from the pixels, but having like the broader context, what it could possibly mean, what is it kind of the broader context for that you could actually do much better. Clearly, PDF has a lot of challenges, including what that formatting is on the PDF file, and you want to how much you want to preserve of that and how much you just want to have the raw text out. |
| 01:27:31 | Philipp | So, yeah, that's that's kind of yeah, it's actually a really amazing how much PDF is. Right. |
| 01:27:37 | Philipp | I mean, we've had a really hard problem. It's a really difficult problem. |
| 01:27:41 | Dion | And, you know, a couple of years ago, Philip and I and a few others were working on a big project called Euro Pet. Yeah. |
| 01:27:48 | Dion | And that had hundreds of thousands of patents that were all in PDF format in many cases that we had to get data out of and restructure. But, you know, if we had the tools today, I think it would be a lot easier. |
| 01:28:00 | Dion | Yeah. |
| 01:28:01 | Dion | Okay. |
| 01:28:02 | Dion | Next question. |
| 01:28:04 | Dion | How many languages does AI voice response include and what are the plans to have all languages? And how about picking up slang and so on. |
| 01:28:12 | Dion | Can it keep up? |
| 01:28:14 | Dion | Okay, so today it supports 55 languages and it can speak in about 140. |
| 01:28:22 | Dion | Okay. So, if you add translation in there it's pretty good. |
| 01:28:27 | Dion | With respect to slang it's been trained on slang and a lot of other information as well. So, it can keep up on that. |
| 01:28:33 | Dion | But you can add glossaries and other things to capture the slang and capture unusual words. So, for example, we're working with one of our media companies that we work with right now where we've written a |

| | | |
|---|---|---|
| | | tool that is doing Euro sports like cycling, and they have Indian speech, I'm sorry, English speech recognition, but it's got European names from Belgium and Norway and Netherlands and Netherlands and Germany and all that that are not common in everyday English. |
| 01:29:03 | Dion | And we're able to set an agent to go and say, right, analyze this bicycle race. Find all the races and cyclists. |
| 01:29:11 | Dion | Find all the commentators, analyze the route of the race and get all the street names and and things that could possibly appear in speech, and make sure they're in the glossary before we run real time transcription, which is then putting the subtitle up on the screen. |
| 01:29:27 | Dion | So, all of those things are possible. You just got to be smart about it. |
| 01:29:32 | Dion | Okay. |
| 01:29:34 | Dion | Somebody is asking about because of AI assisted translation, should they start learning and training and using standard Cat tools? |
| 01:29:43 | Dion | It's more than that. The entire process from front to back is changing everything from data ingestion to how you make glossaries and things like that. |
| 01:29:52 | Dion | I suggest you have a look at the blog post. When it's up there. |
| 01:29:56 | Dion | You'll find a lot of information about that. And it's got a lot of information about what linguists and. |
| 01:30:02 | Dion | Organizations can do as well. |
| 01:30:08 | Dion | Where are we? |
| 01:30:10 | Dion | Okay. |
| 01:30:11 | Dion | Do you know any existing multi-modal translation platforms to date? Sure. |
| 01:30:16 | Dion | ChatGPT right out of the box. Yeah. |
| 01:30:19 | Dion | I mean. |
| 01:30:19 | Philipp | There's also a kind of seamless for MT. That's kind of a major effort to make things kind of more kind of speech and text. |
| 01:30:27 | Philipp | There's like work from, from Google Audio Palm and so on. So, there's various kind of these models and research stages. |
| 01:30:36 | Philipp | Clearly, we never really know what OpenAI is doing with ChatGPT, but they definitely demoed that technology. But we don't really know how much it's kind of chained together or like yeah, it's really end to end process. |

| | | |
|---|---|---|
| 01:30:49 | Philipp | So, this is a really hot topic where, yeah, things are in various stages of development and deployment. |
| 01:30:57 | Dion | Another question. Similar to before, but a different angle. |
| 01:31:00 | Dion | So, does this mean that cat tools are made redundant? |
| 01:31:05 | Dion | The answer is no, not by a long shot. In fact, they're just going to get bigger, better and smarter. |
| 01:31:11 | Dion | They'll be better integrated. They'll talk to more platforms. |
| 01:31:14 | Dion | They will do more things. |
| 01:31:17 | Philipp | Just to give you an example, what you would probably want from a cat tool is that you can just mark up a word and say like, hey, can you give me a different options? Right? This is something that Cat tools can do, but clearly underlying language model can do. |
| 01:31:30 | Philipp | So, there's a lot still work to be done to build really good user interfaces. Exactly. |
| 01:31:35 | Dion | The user interfaces do need a lot of work. |
| 01:31:37 | Dion | Okay. |
| 01:31:39 | Dion | With upcoming regulations for accessibility coming up across countries, how do you see real time translation addressing this? |
| 01:31:46 | Dion | I'm not sure which variation or flavor of accessibility regulations you're talking about, but let me address one of them. |
| 01:31:55 | Dion | So, generating captions and things for people with disabilities, hard of hearing, hard of sight, things like that. There's a lot of work going on and that, including work from our team, so that we can actually generate those captions directly by seeing the background, hearing what's going on, recognizing noises. |
| 01:32:15 | Dion | So, recognizing a car going by or a door slamming or recognizing the tone of somebody's voice so that you can determine that it's angry as opposed to happy, those kinds of things. So, there's a lot of regulations in Europe and other countries or regions saying that these kinds of things have to have a certain percentage of content to be accessible. |
| 01:32:38 | Dion | Okay. Personally, these days I watch every movie with subtitles on when I'm on Netflix or Amazon Prime or whatever, because I just prefer to watch it that way. |
| 01:32:47 | Dion | And I think a lot of people do now, but it's certainly getting very interesting. |
| 01:32:53 | Dion | Okay. |
| 01:32:54 | Dion | So, we have another question here. |

| | | |
|---|---|---|
| 01:32:57 | Dion | Current LM performance quickly degrades as you go down the language list from high resource to low resources. What are your solutions and predictions for this? |
| 01:33:04 | Dion | Very good question. |
| 01:33:06 | Philipp | Yeah I love this question. This is like one of my favorite research topics right now. |
| 01:33:10 | Philipp | And it's actually really astonishing how much kind of the big vendors for kind of these large language models support languages. I mean, if you look at kind of the official prescription description of which languages they support, it's usually maybe a handful, maybe a dozen. |
| 01:33:28 | Philipp | What was it? Llama three had like a list of 12 languages that surprisingly concluded tie. |
| 01:33:33 | Philipp | Right. But it's really, really limited. |
| 01:33:36 | Philipp | I think that is fundamentally going to change in the future and has to change in the future. But it certainly seems to be happening, is that various governments across the world kind of fund large language model development. |
| 01:33:46 | Philipp | So, there's European efforts to build large language models for European languages. |
| 01:33:50 | Philipp | There's a lot going on in India and building languages for all the Indian languages. But I mean, these are kind of fairly targeted solutions. |
| 01:33:59 | Philipp | Clearly, there's challenges to build these models because the vast majority of training data is available in English. And then, you know, the good most of most of the rest is in like a few dozen languages and balancing the training data for the low resource languages or high risk languages. |
| 01:34:14 | Philipp | So, there are technical problems, but ideally you want to have a model that doesn't really matter what language you interact with, you should always get the same responses, and that is currently not the case. So, if you ask the same question, different languages, you get different answers. |
| 01:34:28 | Philipp | But. |
| 01:34:30 | Philipp | It shouldn't be like that. |
| 01:34:32 | Dion | I agree, however, we have seen and Philip, you might know a bit more about this than I do. |
| 01:34:39 | Dion | We have seen that even when you ask a question in, let's say, Arabic, it can leverage the data that's available in English Yeah. And draw down on that. |

| | | |
|---|---|---|
| 01:34:50 | Dion | So, it's not just, oh, only the Arabic data or only that, you know, other language data you've got actually across the board. |
| 01:34:59 | Philipp | Yeah. I think in a webinar in late last year, we actually had an example where I just asked a language model about myself. |
| 01:35:05 | Philipp | Okay. I'm not modest. |
| 01:35:06 | Philipp | And if I ask in German, like which universities I went to, it had a perfect list. But if I asked the same question in Korean. |
| 01:35:13 | Philipp | It was kind of semantically in the right space. All the universities were like German universities. |
| 01:35:17 | Philipp | They were like all computer science related, but like, every single fact was wrong. |
| 01:35:21 | Philipp | Right. But still. |
| 01:35:23 | Philipp | So, there's some knowledge transfer across languages, but not all of it. So, how do you improve? |
| 01:35:29 | Philipp | This is, as I said, this is a research topic I currently work very actively on. Right. |
| 01:35:33 | Dion | Okay. |
| 01:35:34 | Dion | Is there work on using translation memories as rag as, you know, to drive LLM and things around there? Absolutely. |
| 01:35:42 | Dion | We're doing it today. |
| 01:35:44 | Dion | So, we're Extracting terminology from translation memories using language model and other things. We're also using them as examples of how to use terminology in context. |
| 01:35:58 | Dion | Okay. And that's proving very, very fruitful. |
| 01:36:01 | Dion | So, you know, there's some very good examples from that. |
| 01:36:05 | Dion | You know, there's you can draw down on lots of memories. But again you need to make sure that you've got the right content. |
| 01:36:14 | Dion | One of the techniques that we've been using is to create synthetic translation memories or synthetic examples in the target language, so that here's my term and always have an example of how it's used. |
| 01:36:27 | Dion | Right? And even having a description that, you know, let's say it's a a term and it will say, you know, this term like let's say cloud services, it would describe quite literally services hosted by major organizations such as AWS, Google and Amazon or Microsoft in the cloud to offer compute functionality. |

| | | |
|---|---|---|
| 01:36:51 | Dion | So, it would give that definition at time of translation so that the language model knows how to leverage that. And telling it what to do delivers a totally different translation. |
| 01:37:03 | Dion | So, there was a paper a little while back where Google did a simple task. They mirrored all the basic steps of a human translator, and they did it in a pretty primitive model, but it showed really promising results, and we do similar things in our platform. |
| 01:37:18 | Dion | I have to say though, right now it's slow, but still thousands of times faster than a human. But you get much, much better quality when you actually do mirror the steps of a human. |
| 01:37:31 | Dion | And that's a key point. There's no free lunch. |
| 01:37:34 | Dion | So, there's going to be a trade off somewhere. |
| 01:37:37 | Dion | You know, getting the terminology right. You know, it can generate fantastic terminology for you. |
| 01:37:43 | Dion | You've still got to check it, make sure it's good, or you can risk it and it might be okay. And whenever it gets it wrong, just go in and do a tweak. |
| 01:37:50 | Dion | That's another way, right? It's iterative. |
| 01:37:54 | Dion | Okay. |
| 01:37:55 | Dion | That's all we've got time for today. |
| 01:37:57 | Dion | If you do have any more questions, please don't hesitate to email me or Philip. You can contact us at sales at Amazon.com and they'll pass the message on to us. |
| 01:38:08 | Dion | As I mentioned previously, there's a whole series of blog posts that will be published on Monday that cover everything we talked about today in great detail. The slides will be shared, a transcription will be shared, analysis of the transcription will be shared, and of course the video will be shared. |
| 01:38:26 | Dion | So, thank you again for your time today and we look forward to talking to you again soon. |
| 01:38:31 | Dion | Thank you. Have a great day. |
| 01:38:32 | Dion | Bye bye. |