Scaling a state-of-the-art transcription model for high-performant cost-efficient deployment

Table of Contents

tl;dr:

- Simplismart's MLOps platform revolutionises deploying and managing deep-learning models, including the Whisper multi-lingual speech-to-text model developed by OpenAI.
- The platform **reduces latency by 36%**, increases transcription **accuracy by over 8%**, and achieves more than **15 times affordability** (approximately $0.00028/min) compared to existing solutions while making the model **streamable for real-time transcriptions**.
- Simplismart's MLOps platform finds applications across industries, enabling efficient transcription services, customer support, content creation, market research, healthcare, and voice-enabled applications.

> 💡 This TL;DR summary provides a condensed overview of the key points covered in the whitepaper. For a more comprehensive understanding, we recommend referring to the complete document.

Introduction

In the fast-paced world of machine learning, deploying and managing models in production can be daunting. The complexities of MLOps demand robust solutions that are both efficient and cost-effective. In this whitepaper, we introduce Simplismart, an innovative MLOps platform designed to revolutionise the deployment of machine learning models. Specifically, we explore its optimization for the Whisper model, highlighting the substantial improvements in speed, accuracy, and cost-effectiveness that Simplismart brings to the table. By leveraging cutting-edge technologies and streamlining the deployment process, Simplismart offers an out-of-the-box solution that transforms MLOps into a seamless and economical experience.

Background

In recent years, the demand for MLOps solutions has skyrocketed as organisations strive to harness the power of machine learning models in real-world applications. MLOps involves the complex orchestration of various processes, including model training, testing, deployment, monitoring, and maintenance. However, traditional MLOps and ASR solutions often struggle to balance efficiency, accuracy, and cost-effectiveness.

In light of these challenges, Simplismart aims to revolutionise the MLOps landscape by providing an all-in-one platform that is faster, more accurate, and over 15x cheaper than

existing solutions. In the following sections, we delve into the specifics of the Whisper model and how Simplismart optimizes its deployment to achieve these remarkable results.

About Whisper

The Whisper model, developed by OpenAI, is a state-of-the-art multi-lingual speech-to-text model that has revolutionised the field of automatic speech recognition (ASR). Whisper leverages advanced deep-learning architectures and techniques to accurately transcribe speech into written text across multiple languages and dialects.

One of the key strengths of the Whisper model lies in its exceptional accuracy and robustness, even in challenging acoustic environments. Through extensive training on vast amounts of multilingual and multitask supervised data, Whisper has achieved remarkable performance, outperforming many existing ASR systems.

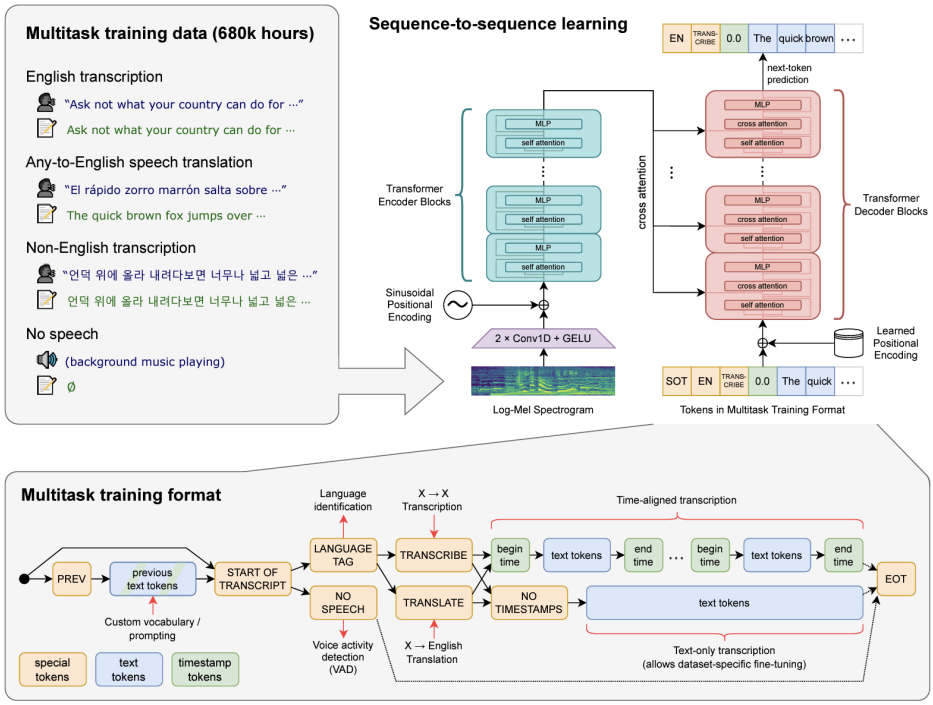

Figure 1: Overview of the Whisper architecture and training approach. (source: Robust Speech Recognition via Large-Scale Weak Supervision)

Whisper's versatility extends beyond its ability to transcribe speech in multiple languages. It excels at handling diverse speech styles, accents, and variations, making it well-suited for applications such as transcription services, voice assistants, call centre analytics, and more. Whether it's conversational speech, professional presentations, or even noisy environments, Whisper demonstrates impressive adaptability and consistently delivers accurate transcriptions.

Existing Solutions

Existing speech-to-text services, while widely used, often have limitations in terms of speed, accuracy, and language support. These services may struggle to accurately transcribe speech in challenging acoustic environments, with varying accents or dialects. Achieving high accuracy with these services can require significant post-processing efforts, impacting the overall efficiency and reliability of the transcription process.

Cost-effectiveness is another area where existing MLOps solutions and ASR services can fall short. Building and maintaining MLOps infrastructure can incur substantial expenses, including hardware costs, licensing fees, and ongoing maintenance. Similarly, utilizing ASR services on a large scale can lead to escalating costs, especially for organizations with significant transcription requirements. These cost considerations can hinder the widespread adoption of ASR technology, particularly for smaller businesses and resource-constrained environments.

To address these challenges, Simplismart's MLOps platform provides a transformative alternative that overcomes the limitations of traditional solutions and ASR services. By optimizing the deployment of the Whisper model, Simplismart enables organizations to achieve superior performance in terms of speed, accuracy, and cost-effectiveness. In the following sections, we explore how Simplismart's platform achieves these remarkable improvements, empowering businesses to deploy and manage deep-learning models, including the Whisper model, with unprecedented ease and efficiency.

Simplismart MLOps Platform

Simplismart takes the complexity out of MLOps with its out-of-the-box platform that streamlines the entire machine learning model development lifecycle. Whether it's the Whisper model or a diverse array of other cutting-edge models, Simplismart's platform offers an unprecedented level of simplicity, efficiency, and cost-effectiveness.

One of the standout features of Simplismart's MLOps platform is its user-friendly interface. With an intuitive design and user-centric approach, Simplismart empowers organizations to effortlessly develop, deploy, and manage various machine learning models without requiring extensive technical expertise. By simplifying the deployment process, Simplismart significantly reduces the time and effort required to put different models into production.

Furthermore, Simplismart's platform incorporates automated mechanisms for optimizing a wide range of models, including the Whisper model. This automated optimization not only saves time but also ensures consistently accurate and reliable predictions from the deployed machine learning models.

In addition to its ease of use and automated optimization capabilities, Simplismart's MLOps platform offers scalability and robustness. Built on a foundation of scalable infrastructure and distributed computing, the platform can efficiently handle large volumes of data and high-velocity data streams for a variety of models. This scalability enables real-time deployment of different models in demanding production environments without compromising performance or reliability.

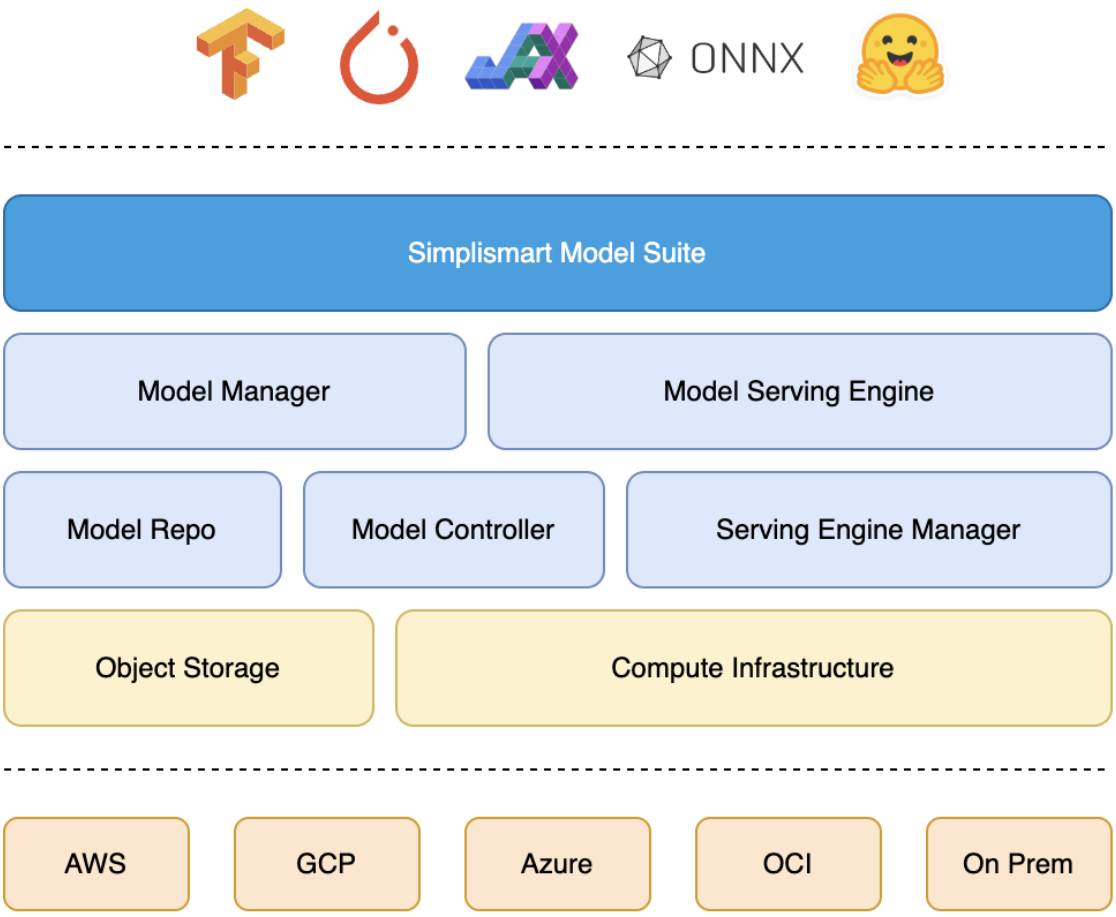

Figure 2: Overview of the Simplismart Model Suite architecture

Moreover, Simplismart's MLOps platform introduces a remarkable level of cost-effectiveness. By leveraging cloud-native technologies and optimized resource allocation, Simplismart significantly reduces the infrastructure costs associated with deploying and maintaining various machine learning models. This cost optimization empowers organizations, regardless of their size or budget, to fully harness the potential of advanced models without being burdened by excessive expenses.

Enhanced Speed, Accuracy, and Cost-effectiveness for the Whisper Model:

Simplismart's MLOps platform drives remarkable enhancements for the Whisper model, delivering outstanding speed, accuracy, and cost-effectiveness. Extensive testing, on over a thousand audio samples per language, has shown that Simplismart's platform surpasses existing solutions, providing substantial benefits for organisations leveraging ASR models.

Cost-Effectiveness

Cost-effectiveness is a significant advantage of Simplismart's MLOps platform, making it more than **15 times affordable** than then next most affordable provider for ASR. By leveraging model-level optimisations and optimized resource allocation, Simplismart minimizes infrastructure costs associated with model deployment and maintenance. This cost optimization empowers organizations to leverage the powerful speech-to-text capabilities of the Whisper model without incurring excessive expenses, making it an affordable solution for a wide range of use cases and budgets.

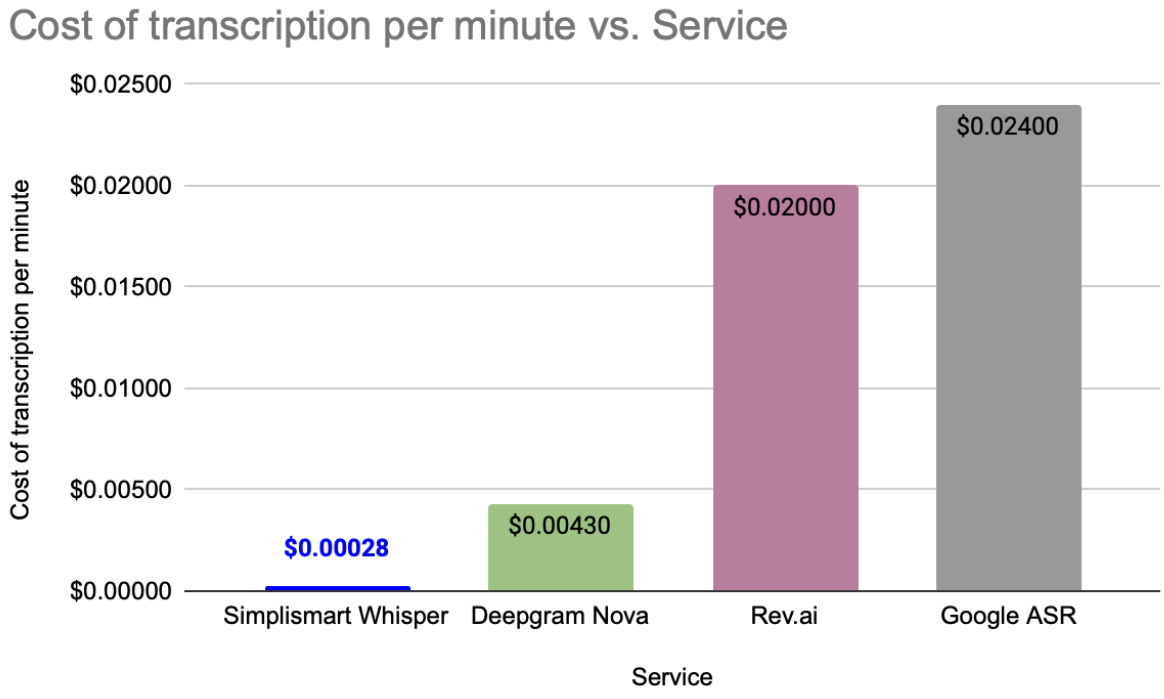

Figure 3: A bar chart of the cost of transcription per minute of audio across providers. The lower the cost, the better.

Cost per minute calculation

To determine the cost per minute for audio transcription using the given data, we can calculate the cost per audio file and then convert it to a cost per minute.

1. Calculate the cost per audio file:
    - Each machine can transcribe 2 audio files parallelly
    - Each transcription of an audio of length 120 Seconds takes 8 seconds.
    - This means each audio file takes 4 seconds (8 seconds / 2).
    - Given that a T4 GPU machine with 8 CPUs costs $0.51 per hour, we need to convert the cost to per-second basis:
        - Cost per second = $0.51 / (60 minutes * 60 seconds) = $0.0001833 per second.
    - Multiply the cost per second by the transcription time per audio file:
        - Cost per audio file = $0.0001416 per second * 4 seconds = $0.0005667.
2. Convert the cost per audio file to cost per minute:
    - Since each audio file's length is 2 minutes, the cost to transcribe a minute of audio:
        - Cost per minute of audio = $0.0005667 per audio file / 2 minutes per audio file = $0.0002834 / minute

Therefore, the **cost per minute** for audio transcription is approximately **$0.0002834**.

> 💡 Please note that the above calculation is based on the provided data and assumptions. Additional factors, such as any additional costs/discounts or specific pricing structures, should be considered for a more accurate cost estimation, such as discounts on reserved and preemptible instances.

Accuracy

In terms of accuracy, Simplismart's MLOps platform achieves more than a **8% increase in transcription accuracy** compared to existing solutions. Through optimization techniques such as language specific fine-tuning, the model delivers highly precise speech-to-text transcriptions. This improvement in accuracy ensures reliable and high-quality transcriptions, even in challenging acoustic environments and across multiple languages.

	English	Spanish	Hindi
Simplismart	**3.84%**	**5.66%**	**13.90%**
Deepgram	5.21%	18.10%	34.55%
Rev.ai	5.21%	6.37%	24.20%
Google ASR	6.05%	6.21%	19.23%

Table 1: Word-Error-Rate (WER) across providers and languages on over a thousand samples per language sampled from a mix of Mozilla Common Voice and Google FLUERS data. The lower the WER the better.

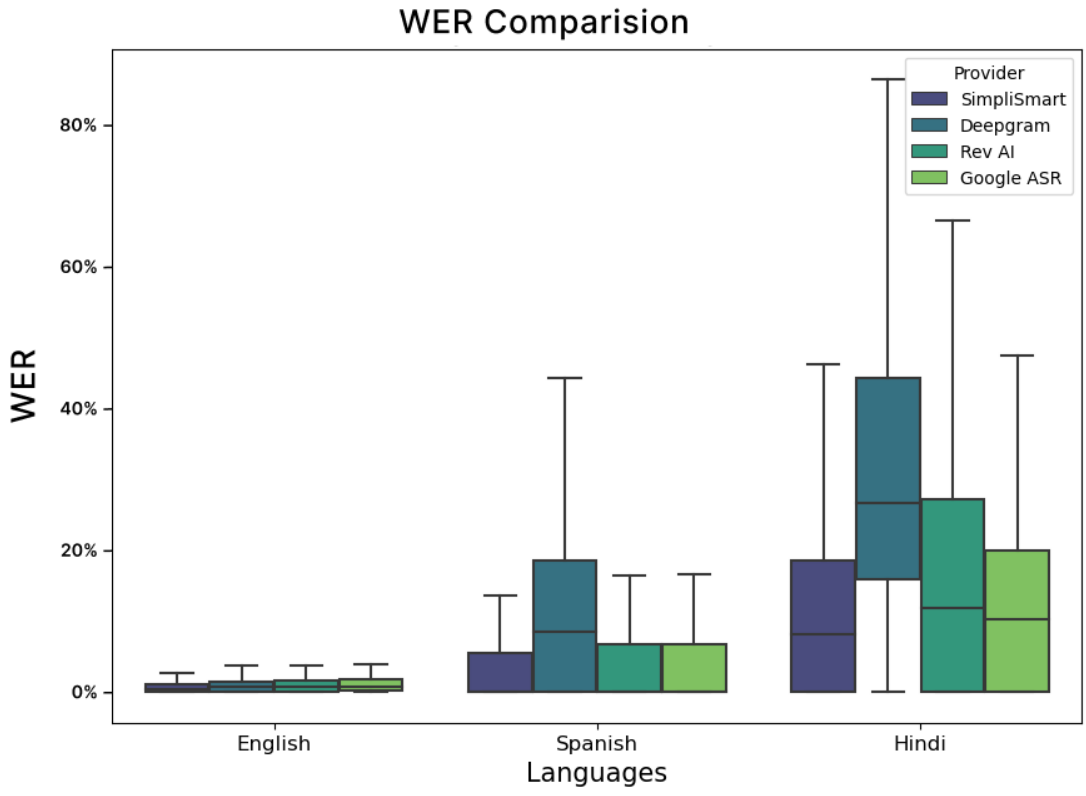

Figure 4: Word-Error-Rate (WER) across providers and languages on over a thousand samples per language sampled from a mix of Mozilla Common Voice and Google FLUERS data. The lower the WER the better.

Speed

In terms of speed, Simplismart's MLOps platform achieves a remarkable **36% average reduction in latency** compared to existing solutions. By leveraging optimized infrastructure and efficient resource allocation, Simplismart ensures that organizations can obtain real-time transcriptions with minimal delay, including streaming audio and transcriptions in real-time. This reduction in latency enables organizations to process audio data rapidly, facilitating timely decision-making and enabling efficient workflows.

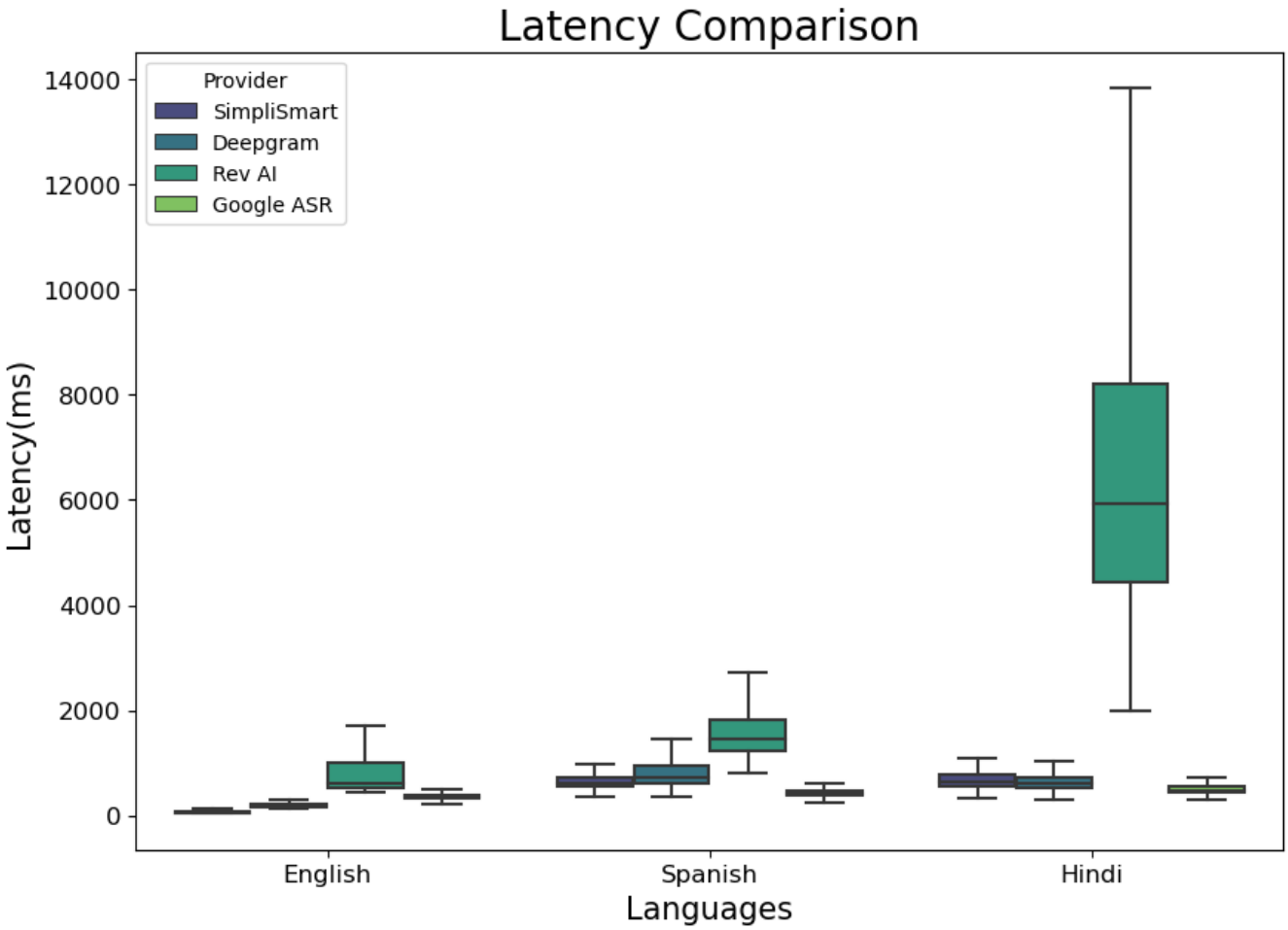

Figure 5: A box plot of latency (ms) per second of audio across providers and languages on over a thousand samples per language sampled from a mix of Mozilla Common Voice and Google FLUERS data. The lower the latency, the better.

With Simplismart's MLOps platform, organizations can benefit from enhanced speed, accuracy, and cost-effectiveness when deploying and managing the Whisper model. By reducing latency, improving accuracy, and optimizing costs, Simplismart empowers organizations to extract valuable insights from audio data with unparalleled efficiency.

Wide Range of Use-Cases for the Whisper Model:

The Whisper model, deployed and managed through Simplismart's MLOps platform, finds extensive applicability across various industries and use-cases. Its superior speech-to-text capabilities and optimized performance make it an invaluable asset for organizations seeking accurate and efficient audio data processing. Here are some prominent use-cases where the Whisper model, powered by Simplismart's platform, excels:

1. **Transcription Services:** The Whisper model's exceptional accuracy and speed make it an ideal solution for transcription services. Simplismart's MLOps platform enables transcription service providers to process audio files quickly and accurately, delivering high-quality transcriptions in record time. From academic institutions to media agencies, the Whisper model offers a cost-effective and efficient solution for transcribing interviews, lectures, podcasts, and more.
2. **Customer Support and Call Centers:** Streamlining customer support and call center operations is essential for businesses. The Whisper model, integrated with Simplismart's MLOps platform, enables automatic transcription of customer calls, facilitating real-time analysis and sentiment analysis. This application enhances customer service quality, aids in training customer support representatives, and provides valuable insights for improving overall customer experience.
3. **Sales enablement and notetaker tools:** Transcribing voice calls over meeting platforms like zoom, google meets is crucial and can be often challenging. Handling unit economics with specific subscriptions becomes difficult most of the times. Moreover, enterprises generally prefer to keep their data only on their premises and are not comfortable giving it to a third party API. All of these problems can be solved by deploying Simplismart’s serving suite in-house at a much cheaper cost and providing a better accuracy.
4. **Content Creation and Media:** Content creators, including podcasters, vloggers, and journalists, can benefit from the Whisper model's accurate and fast speech-to-text transcription capabilities. By using Simplismart's MLOps platform, content creators can easily generate transcriptions of their audio content, allowing for efficient content indexing, searchability, and repurposing. This simplifies content editing, enables captioning for accessibility, and facilitates the creation of written content from audio interviews or discussions.
5. **Market Research and Data Analysis:** The Whisper model's ability to transcribe audio data efficiently enables organizations to derive valuable insights from market research interviews, focus groups, and surveys. Simplismart's MLOps platform ensures accurate and timely transcription, accelerating the analysis process and enabling organizations to

extract key information for strategic decision-making, product development, and market insights.

6. **Healthcare and Medical Transcription:** Simplismart's MLOps platform, and its optimisations for the Whisper model, offers significant advantages in the healthcare industry. It simplifies medical transcription, enabling healthcare providers to convert patient records, doctor-patient interactions, and medical dictations into accurate and accessible text format. This streamlines administrative processes, improves documentation accuracy, and enhances patient care.
7. **Voice Assistants and Virtual Agents:** Voice-enabled applications, such as voice assistants and virtual agents, can leverage the Whisper model's superior speech-to-text capabilities. Simplismart's MLOps platform enables organizations to integrate the Whisper model seamlessly, enhancing the accuracy and responsiveness of voice-driven applications, improving user experiences, and expanding the range of voice-enabled functionalities.

These are just a few examples of the diverse use-cases where Simplismart's MLOps platform, and its optimisations for the Whisper model, can deliver significant value. Its accurate and efficient speech-to-text capabilities open up new possibilities for organizations across industries, improving productivity, enabling data-driven decision-making, and enhancing user experiences

Conclusion: Unlocking the Power of the Whisper Model with Simplismart's MLOps Platform

Simplismart's MLOps platform offers unprecedented speed, accuracy, and cost-effectiveness for deploying and managing the Whisper model. With reduced latency, over 8% increase in accuracy, and more than 15 times cost savings compared to existing solutions, organizations can harness the full potential of the Whisper model for diverse applications. Simplismart's platform transforms audio data into valuable insights, accelerating innovation and driving data-driven success across industries.

Embrace the transformative capabilities of Simplismart's MLOps platform and unlock the true potential of deep learning in your organization. Experience faster, more accurate, and cost-effective speech-to-text transcription that propels you ahead in the era of data-driven insights and intelligence.

To learn more about Simplismart's MLOps platform and how it can revolutionize your organization's deployment and management of your deep learning models, visit our website or contact our team at contact@simplismart.ai. Start your journey towards enhanced efficiency, productivity, and data-driven success today.